Surveillance Methods for Monitoring HIV Incidence and Drug Resistance

Surveillance Methods for Monitoring HIV Incidence and Drug Resistance

A dissertation presented

by

Natalie Mae Exner

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

March 2014

Dissertation Advisor: Professor Marcello Pagano Natalie Mae Exner

Surveillance Methods for Monitoring HIV Incidence and Drug Resistance

Abstract

Disease surveillance is the continuous collection, analysis, and interpretation of health-related data. Information gained from routine HIV disease surveillance is vital to national program managers deciding to implement new prevention or treatment programs. In this dissertation, we describe methods for estimation of HIV incidence and the prevalence of HIV drug resistance.

HIV incidence estimation is critical for identifying at-risk populations for targeted interventions and measuring the effectiveness of these interventions. We provide an in-depth literature review of the available options for estimating HIV incidence, including cross-sectional assays. Next, we describe a novel cross-sectional assay for HIV incidence estimation that discriminates between recent and long-term infections on the basis of within-host viral diversity. Diversity is measured using a version of Shannons entropy that we adapt to improve discriminatory ability. These adaptations include breaking the within-host sequence alignment into smaller sections to allow for more nuanced detection of within-host variability, and we suggest an algorithm for adjusting for multiple HIV infections using clustering methods.

HIV drug resistance surveillance guides national programmatic managers identifying effective treatment regimens for HIV-infected individuals in their countries. We describe a large-scale consulting project with the World Health Organization to redesign their guidance for pre-treatment and acquired HIV drug resistance surveillance in low- and middle-income countries. Our consulting work prompted a variety of

interesting statistical questions that we address in a series of papers. We describe a novel method for calculating sample sizes for two-stage clustered surveys in which the finite population correction can be applied. Our method results in a sometimes dramatic decrease in sample size while still achieving the desired precision. We introduce a novel acquired HIV drug resistance outcome for measuring viral load suppression that incorporates information on patient loss-to-follow-up. This outcome has increased epidemiological utility over previously used outcomes. Finally, we evaluate methods for confidence interval estimation for proportions measured in surveys and provide recommendations for their use.

Contents

List of Figures

List of Tables

To my best friend and future husband, Ethan.

Acknowledgments

I would like to take this time to briefly thank those who have been most influential in my academic career.

To my advisor, Marcello Pagano, thank you for the outstanding mentorship. It has been a pleasure working with you and learning from you. I would also like to acknowledge Victor De Gruttola and Vladimir Novitsky for their contributions to this dissertation and for the knowledge that they shared with me. From the WHO HIV Drug Resistance Team, I thank Silvia Bertagnolio, Jhoney Barcarolo, and Michael Jordan, and I look forward to our continued collaboration.

To my classmates, Shira Mitchell, Mark Meyer, and Stacey Ackerman-Alexeeff, thank you for your companionship throughout this process. To my closest friends, Megan Scarborough, Sarah Taylor, Stephanie Santarpio, thank you for keeping the fun in these past few years. To my parents and sister, thank you for your incredible love and support. To Ethan Dean, my future husband, none of this would have been possible without you.

This work was supported by NIH grants T32AI07358 and RO1AI097015.

1. Introduction

My doctoral dissertation focuses on two important aspects of HIV-1 disease surveillance: incidence estimation and monitoring of drug resistance. Disease surveillance is the continuous collection, analysis, and interpretation of health-related data, and disease surveillance is critical for public health practice. Information gained from surveillance is vital to national public health managers deciding to implement new prevention or treatment programs. Because HIV/AIDS disproportionately affects under-served populations, logistical simplicity and low cost are very important factors in evaluating these surveillance methods. Surveillance must be sufficiently feasible to be regularly executed in low- and middle-income countries. Otherwise, countries may not implement these methods and will lack the data necessary to make informed programmatic decisions.

In the Chapters 2 and 3, we describe our methods for estimation of HIV-1 incidence. Incidence measures the rate of recent infections, and it is critical for understanding transmission dynamics, identifying at-risk populations for targeted interventions, measuring the effectiveness of community-level interventions, and calculating sample size requirements for randomized trials. In Chapter 2, we provide an in-depth literature review of available options for estimating HIV incidence, emphasizing the serious challenges in this field. Standard follow-up of an HIV-negative cohort is prohibitively expensive, logistically complicated, and subject to biases. Other methods that involve repeated prevalence surveys require a variety of assumptions and are only applicable in generalized epidemics. Ultimately, cross-sectional surveys that discriminate between recent and long-term infections using a host or viral marker are most appealing. The majority of existing assays measure immunological factors, but these assays have drawbacks including subtype variability and individuals who never test as long-term on the assay despite many years of HIV infection (long-term non-progressors). Diversity-based assays are an emerging alternative; they rely on the immense increase in viral diversity that occurs within a host over the course of an infection.

In Chapter 3, we describe a novel assay for discriminating between recent and long-term infections using viral diversity. We adapt an existing diversity measure known as Shan-

non's entropy. Entropy is a concept from information theory, and it is used to measure the variability in a multinomial outcome. After a person is infected with HIV, the virus mutates resulting in a diverse range of sequence patterns within a single individual. Here, the multinomial outcome of interest is each unique viral sequence pattern within an individual. In our paper, we demonstrate that by dividing the sequence alignment into smaller sections before calculating entropy, we increase the discriminatory ability of the assay as measured by the area under the receiver operator characteristic curve. In addition, we introduce a method for adjusting for the presence of multiple infections. Multiple infections confound the relationship between time since infection and diversity because individuals recently infected with multiple viruses can have high diversity. We use clustering methods to separate viral lineages and measure within-lineage entropy rather than overall entropy.

In the Chapter 4, we describe a large-scale consulting project with the World Health Organization (WHO) to redesign their guidance for HIV drug resistance surveillance in low- and middle-income countries. This guidance for surveillance in patients initiating treatment (pre-treatment resistance) and patients on treatment (acquired drug resistance) will be adapted by countries into fully realized survey protocols. These protocols will then be implemented by the countries with limited support from the WHO. Thus, simplicity and feasibility are critical for determining the success of this guidance. Nonetheless, we emphasize the importance of collecting representative data and analyzing the data in a statistically rigorous fashion. In a report included in this dissertation, we describe the surveys and the statistical decisions made during the consulting process, focusing primarily on sample size calculations and survey analysis. Guidance for the two surveys is currently being published by the WHO.

Our consulting work prompted a variety of interesting statistical questions which we address in a series of papers. In Chapter 5, we describe a novel method for calculating sample sizes for two-stage clustered surveys in which the finite population correction will be applied. We demonstrate dramatic decreases in sample size requirements while

still achieving the desired precision. This paper has important implications for disease surveillance in small countries or countries with concentrated epidemics. Using previously available methods, sample size requirements can be larger than the total eligible population size in some countries. As a result, these countries do not feel able to properly implement the survey and may not conduct the survey at all. Our method empowers small countries to design drug resistance surveys that will work in their unique setting. The sample size procedure introduced in the paper is already being implemented as a core element of the WHO's HIV drug resistance surveillance guidance.

In Chapter 6, we introduce a novel outcome for the acquired drug resistance survey. The acquired drug resistance survey includes a cross-sectional assessment of viral load in patients on antiretroviral therapy for 12 ± 3 months. As this measure excludes patients who have died or been lost to follow-up, it is biased relative to the population-level prevalence of viral load suppression. As a result, this measure can be misleading in the absence of representative data on patient retention. A country may have high viral load suppression among retained patients, but they may have many patients lost to follow-up. If that country implemented a program to improve patient retention, they may actually observe a decrease in viral load suppression among retained patients because they have captured some of the sickest patients who would have otherwise been lost. Through consultation with the WHO's HIV drug resistance steering group, we constructed a new measure of population-level viral load suppression that assumes that all patients who are lost to follow-up are not virally suppressed. The outcome can be interpreted as a lower bound of viral load suppression, and it has improved epidemiological utility over the unadjusted outcome. In the paper, we derive the properties of this outcome, including point and standard error estimators. We also derive the expected precision of this outcome under a set of assumptions and assess sensitivity to these assumptions through a series of simulations.

In Chapter 7, we evaluate methods for confidence interval estimation for proportions measured in surveys. Extensive literature has demonstrated that the Wald interval per-

forms poorly in the independent and identically distributed setting. Alternative methods, including the Wilson, Jeffreys, and Agresti-Coull intervals, perform much better. The Clopper-Pearson, or 'exact,' interval is another popular option, although it tends to be unnecessarily wide and conservative. There is very little peer-reviewed evaluation of these intervals in survey settings. In our paper, we describe seven methods for calculating confidence intervals for proportions, including adaptions of the Jeffreys and Agresti-Coull intervals that have never been applied in the existing literature. For each method, we describe an approach that does and does not incorporate the design degrees of freedom into the calculations, suggesting a framework to increase logical consistency across formulations. We evaluate these methods via an extensive simulation study. We demonstrate the importance of adjusting for the design degrees of freedom, and we show that the modified Jeffreys and Wilson intervals perform best in terms of confidence interval length and coverage. Finally, we address the topic of truncation, in which the effective sample size from a clustered survey is not allowed to exceed the actual survey sample size. We describe a discrepancy in the existing literature on truncation and provide our recommendations for this setting.

2. The challenges of HIV incidence estimation

Natalie Exner

2.1 Background

The ability to accurately estimate incidence, or the risk of acquiring infection within a given period of time, is a critical component of any human immunodeficiency virus type 1 (HIV-1) disease surveillance program. A reliable estimate of incidence in a given population allows investigators to understand transmission dynamics, evaluate the performance of prevention programs, and identify high-risk populations so as to efficiently utilize resources (Rutherford et al., 2000). Accurate incidence estimates are also critical when calculating sample size requirements during the design phase of prevention trials. Despite the importance of this indicator, it is extremely difficult to estimate HIV-1 incidence in practice (Family Health International, 2009). Methods currently available include direct observation of a seronegative cohort, back-calculation methods, modeling the epidemic in the general population using serial prevalence surveys, and cross-sectional assays which rely on the evolution of host or viral markers. All existing methods have serious limitations, and the search for an estimator which is both unbiased and logistically feasible continues.

Direct observation of a seronegative cohort was once considered the 'gold standard' for the measurement of HIV-1 incidence. In practice, this method is rarely used because it requires enormous sample sizes, making it both logistically complicated, prohibitively expensive, and unsustainable even in resource-rich settings (Family Health International, 2009). Furthermore, there is potential for selection and follow-up bias because those who are willing to enroll in such a trial and who are not lost to follow-up may not be representative of the general population. Brookmeyer et al. calculated two estimates of incidence for two STD clinics in India (Brookmeyer et al., 1995). The cohort-based estimate was markedly less than the combined cohort/cross-sectional estimate, and the difference in these estimates was supported by significant behavioral differences between those who returned for follow-up and those who did not. Since individuals enrolled in a prevention trial are counseled on risk reduction, it is reasonable to expect their HIV-1 incidence to

be lower than that of the general population. Thus, direct observation of a seronegative cohort is not an appropriate 'gold standard' for the estimation of HIV-1 incidence.

Until the mid-1990s, incidence could be back-calculated from AIDS surveillance data using an estimate of the distribution of the viral incubation period (Brookmeyer, 1991). Back-calculation methods from AIDS diagnoses require reliable disease surveillance data as well as an accurately described incubation period (Gail and Brookmeyer, 1988). Because of the widespread use of highly active antiretroviral treatment (HAART), this method is no longer in use, but some researchers use back-calculation from other events besides AIDS diagnoses. Extended back-calculation models have been described using a dichotomous measure of disease severity at time of HIV diagnosis (Hall et al., 2008), CD4 cell count (Satten and Longini, 1994; Taffe et al., 2008), and the joint distribution of two HIV antigens (Sommen et al., 2010). These methods require many assumptions, most notably regarding testing behavior since they back-calculate from time of first positive antibody test (Karon et al., 2008). Because of the large amount of individual level variability in CD4 and HIV antigen trajectories as well as the necessity for assumptions about homogeneity of testing behavior across individuals and time, these back-calculation methods are not ideal for the estimation of HIV-1 incidence.

An alternative method infers incidence from changes in age-specific seroprevalence measured by serial surveys. The method works by dividing up the general population into age cohorts so that these cohorts can be tracked through time across repeated surveys; given the prevalence at the time of the last survey, the method accounts for deaths using assumptions about age-specific mortality, and then it calculates the number of new infections that must have occurred in each cohort to achieve the observed prevalence in the current survey (Hallett et al., 2008). The method can only be applied to generalized epidemics in which large-scale demographic surveys are routinely carried out, and it cannot be used to calculate incidence in sub-epidemics, such as among injection drug users (Rehle et al., 2010). The accuracy of the estimate depends on the reliability of the seroprevalence surveys, which can be biased if they do not explicitly account for differential

testing refusal, underrepresentation of mobile groups (Marston et al., 2008), and internal migration (Hallett et al., 2008). Estimation of age-specific mortality is further complicated by the roll-out of HAART and requires additional assumptions regarding age-specific treatment initiation and the effect of treatment on survival (Rehle et al., 2010). Computer packages, such as the Spectrum package (Stover, 2004; Stover et al., 2010), have been developed to model age-specific incidence, but their results are only as good as the seroprevalence data and assumptions on which they rely (Ghys et al., 2004). Thus, while these models may serve as useful tools for understanding transmission dynamics, their usefulness for accurately estimating incidence is limited (Sakarovitch et al., 2007).

2.2 Cross-sectional assays

Cross-sectional assays for recent HIV-1 infection are the most promising methods for incidence estimation; they are cheaper, simpler, and prone to less bias than direct observation of a cohort, and they require many fewer assumptions than back-calculation methods or computer models. These assays distinguish between recent and long-term infections on the basis of a host or viral marker; when the marker is below some threshold, the infection is classified as 'recent', and when the marker is above that threshold, the infection is classified as 'long-term.' Incidence can be calculated from a cross-sectional survey as $I = P/\omega$ where P is the prevalence of recent infections within the at-risk population (i.e. excluding long-term infections), and ω is the mean duration of the period when an infection is classified as 'recent', also called the window period (Rothman et al., 2008). This estimator assumes constant incidence during the period before the survey at least as long as the maximum plausible window period (Brookmeyer and Quinn, 1995). It also requires an accurate estimate of the mean window period in the population of interest, and, as a result, makes the implicit assumption that all individuals with HIV infection have markers that will eventually cross the 'recent/long-term' threshold (Wang and Lagakos, 2009). Most existing assays rely on markers of evolution of the host immune system in

response to HIV infection, though recent papers describe novel assays which exploit viral characteristics.

The first cross-sectional assay for recent infection was described in 1995 by Brookmeyer and Quinn (1995). The assay relies on the presence of detectable p24 antigenemia during the period before seroconversion. The survey first identifies and excludes all antibody positive individuals, and then all antibody negative individuals are tested for p24. This assay is not in use for two primary reasons: 1) the window period for p24 antigenemia is very short (approximately 22.5 days from the first report), and 2) the assay is used to test antibody negative individuals. Even in a high prevalence setting, this assay will require enormous sample sizes to detect enough recent infections for a precise estimate of incidence. The assay also has low sensitivity among individuals in the process of seroconversion (antibody indeterminate) (Hecht et al., 2002). A similar assay was developed using detectable HIV-1 RNA during the pre-seroconversion period. While this assay has better sensitivity and a slightly longer mean window period (Le Vu et al., 2009), it still requires the testing of large samples of antibody negative individuals. Some authors have proposed pooling algorithms for RNA testing to reduce costs (Brookmeyer, 1999; Quinn et al., 2000), but, even with these cost-saving measures, the assay remains much more expensive than an assay which is conducted using only antibody positive samples.

The first test intended for use in seropositive individuals was described by Janssen et al. (1998). By varying the laboratory procedure for the traditional antibody assay (higher dilution, shorter incubation time, higher cutoff), they created a less-sensitive, or detuned, assay, subsequently referred to as the Serological Testing Algorithm for Recent HIV Seroconversion (STARHS). As antibody levels increase during the early stages of infection, seroconversion on the sensitive assay precedes seroconversion on the less-sensitive assay, and the time between these two events is described by the mean window period (129 days (95% CI: 109-149 days) in the initial report of a subtype B cohort (Janssen et al., 1998)). While this assay has significant advantages over p24 and RNA testing in terms of cost and feasibility, additional studies quickly identified its numerous limitations; the assay suffers

from low internal consistency as demonstrated by high coefficients of variation (>20%), most likely because of the magnitude of the dilution (1:20,000) (Kothe et al., 2003); the mean window period varies by HIV-1 subtype (Parekh et al., 2001; Wilson et al., 2004; Young et al., 2003); the assay can misclassify those with advanced infection (Guy et al., 2005; Parekh et al., 2001; Rawal et al., 2003; Wilson et al., 2004), those on antiretroviral treatment (ART) (Killian et al., 2006; Rawal et al., 2003), and 'elite suppressors' (Laeyendecker et al., 2008); and there is evidence of individuals whose markers never cross the less-sensitive threshold regardless of the length of follow-up (Young et al., 2003). As a result, STARHS has low specificity which can lead to greatly overestimated measures of incidence.

With the specific goal of addressing the subtype variability observed for STARHS, the CDC developed the BED capture enzyme immunoassay (BED-CEIA) using gp41 sequences from HIV-1 subtypes B, E, and D (Parekh et al., 2002). The assay measures the ratio of HIV-specific immunoglobulin G (IgG) to total IgG, a proportion which generally increases during the first two years after seroconversion (Parekh and McDougal, 2001). While this assay has advantages over STARHS, including higher internal consistency (Dobbs et al., 2004), lower cost, and commercial availability, a variety of studies have demonstrated that the BED-CEIA has many of the same weaknesses as its predecessor, including a non-zero fraction of the population who persistently test as 'recent' despite demonstrated long-term infection (Hargrove et al., 2008; Karita et al., 2007), misclassification of those with advanced infection (Karon et al., 2008; Marinda et al., 2010), and misclassification of those on ART (Marinda et al., 2010). Furthermore, the mean window period varies widely by subtype, either because of viral or host immunological factors (Karita et al., 2007; Parekh et al., 2011). As a result, in 2005 UNAIDS recommended that the BED-CEIA not be used in routine HIV-1 surveillance, including absolute incidence estimation or monitoring trends (UNAIDS Reference Group on Estimates Modeling and Projections, 2005).

Because of its poor specificity, using the BED-CEIA in a high prevalence setting can lead to

significant overestimation of incidence (Sakarovitch et al., 2007). A variety of estimators have been proposed which adjust for imperfect sensitivity, short-term specificity (relating to infections of duration longer than the window period ω but less than 2ω), and long-term specificity (relating to infections of duration greater than 2ω). The first adjusted estimator, proposed by McDougal et al. (2006), required calibration of four parameters: the mean window period ω, sensitivity σ, short-term specificity ρ_1, and long-term specificity ρ_2. This formula was further simplified by Hargrove et al. (2008) assuming that $\sigma = \rho_1$, yielding only three parameters for calibration. McWalter and Welte (2008) constructed an estimator that makes fewer assumptions than either McDougal or Hargrove and requires calibration of only two parameters: the mean window period ω and long-term specificity ρ_2 (McWalter and Welte, 2009). Given the proportion of seropositive individuals that register under the assay threshold P_0, the true proportion of recent infections is $P_T = \frac{P_0+\rho_2-1}{\rho_2}$. This adjusted prevalence can then be plugged into the traditional incidence formula $I = P_T/\omega$. Their estimator coincides with the maximum likelihood estimator derived by Wang and Lagakos (2009).

Despite the mathematical justification for the McWalter/Welte estimator, the use of adjustments for cross-sectional incidence estimates is hotly debated (i.e. which one, if any) (Brookmeyer, 2009a,b; Hargrove, 2009; McDougal, 2009; Welte et al., 2009). Following the 2005 UNAIDS statement discouraging use of the BED-CEIA, the Global AIDS Coordinator stated that the assay may be used as long as appropriate adjustments are made (Office of Global AIDS Coordinator, 2006). In practice, these adjustments rely on accurate, locally-derived estimates of the calibration parameters (Barnighäusen et al., 2008, 2010; Kim et al., 2010; World Health Organization, 2009b), and if these calibration parameters vary significantly between populations or across time, it is unlikely that the BED-CEIA will be of practical use (Hallett et al., 2009; Welte et al., 2010). One alternative proposed by Wang and Lagakos is an augmented cross-sectional design which longitudinally follows individuals who test as recent on the BED-CEIA with the goal of calculating an internal estimate of long-term specificity; this method has the logistical challenges associated with

individual follow-up, though to a much lesser extent than a complete cohort study, and may be difficult to implement in practice (Wang and Lagakos, 2010). Overall, despite the initial promise of the BED-CEIA, extensive research has demonstrated that it is difficult to obtain accurate estimates of HIV-1 incidence using this assay alone.

Besides the STARHS and BED-CEIA, a variety of other assays have been developed to exploit maturation of the host immune system during HIV-1 infection. Among the most commonly used is the Avidity Index developed in 2003 (Suligoi et al., 2003). While the original paper claimed that the assay was robust to ART use and the presence of advanced disease, subsequent studies have disproved these claims (Chawla et al., 2007; Sakarovitch et al., 2007; Selleri et al., 2007). The IDE-V3 assay developed in 2005 measures antibodies specific to four HIV-1 antigens, including the immunodominant epitope of gp41 (IDE) and V3 peptides (Barin et al., 2005). This assay, also referred to as the enzyme immunoassay for recent infection (EIA-RI), is not appropriate for individuals initiated on ART during early infection (Barin et al., 2005) and has overall low sensitivity (Le Vu et al., 2009; Sakarovitch et al., 2007); there is also preliminary evidence of subtype variability (Le Vu et al., 2009). Other assays include the anti-p24 IgG3 assay (Wilson et al., 2004), the line immunoassay (Schüpbach et al., 2007), the particle agglutination assay (Hong et al., 2007), and a new multi-subtype avidity-based assay developed by the CDC (Wei et al., 2010). These assays are summarized in a review paper by the WHO Working Group on HIV Incidence Assays (Busch et al., 2010). Overall, I believe that future evaluation of these immunoassays will reveal many of the same drawbacks as the existing immunoassays, including low specificity and subtype variability. While immunoassays provide important information regarding duration of infection, I do not think that they alone can provide accurate and reliable estimates of HIV-1 incidence.

2.3 Diversity-based assays

An emerging alternative to immunoassays are diversity-based assays which exploit the increase in viral diversity that occurs during the early stages of HIV-1 infection. The majority of infections are seeded by a single founder virus (Keele, 2010), resulting in an initial viral population with zero diversity. Because of the high error rate of the reverse transcriptase enzyme (Ji and Loeb, 1994) and the rapid turnover rate *in vivo* (Ho et al., 1995; Wei et al., 1995), the virus is able to rapidly diversify within its new host (Coffin, 1995). There is strong evidence that viral genetic diversity increases linearly with time during this initial phase of HIV-1 infection (Frost et al., 2005; Kearney et al., 2009; Lee et al., 2008, 2009; Shankarappa et al., 1999). This accumulation of mutations has also been described using a Poisson distribution (Keele et al., 2008; Lee et al., 2009; Leitner and Albert, 1999). The envelope (*env*) gene, which codes for the viral surface proteins that interact with the host immune system, has the greatest potential for diversity within the HIV-1 genome. Variability in *env* is adaptive in that it helps the virus evade host immune pressures (Holmes et al., 1992). In fact, unlike other regions of the genome, there is compelling evidence that *env* is under positive, or diversifying, selection, promoting variability in the host viral population rather than selecting for only the 'fittest' quasispecies (Bonhoeffer et al., 1995; Ganeshan et al., 1997; Overbaugh and Bangham, 2001; Piantadosi et al., 2009; Poss et al., 1998; Yamaguchi and Gojobori, 1997). Overall, diversity-based assays of recent infection are promising because of the immense within-host variability of the HIV-1 genome.

A variety of diversity-based assays have been described in the recent literature. There is one which uses the fraction of ambiguous nucleotide calls obtained during bulk sequencing of the *pol* gene as a proxy for overall diversity (Kouyos et al., 2011; Wilson et al., 2011; Andersson et al., 2013). Using a Swiss subtype B cohort, they found evidence that this fraction increases linearly during the first eight years of infection (Kouyos et al., 2011). The advantage of this type of test is that it is easily implementable as bulk sequencing of

this region is already part of standard genotypic resistance testing procedures. It also has certain disadvantages: nucleotide ambiguity is inherently a binary measure and may not adequately describe the diversity present at a particular position; the lower limit of detection is typically 20% (Kouyos et al., 2011), meaning the assay will underestimate diversity if there are minority quasispecies in the population; and, while diversity in *pol* correlates with diversity in other regions, genetic bottlenecking could occur in response to selective pressures, thereby reducing specificity. Another recent assay suggests measuring diversity in multiple regions of the genome, including *env*, *pol*, and *gag*, to increase robustness against such genetic bottlenecking (Cousins et al., 2011). This assay uses high resolution melting (HRM), which measures diversity without sequencing using the melting characteristics of DNA duplexes (Towler et al., 2010). Again, the advantage of this assay is its implementability, being in a 96-well plate format and taking only a few minutes to run. A disadvantage not addressed is the likely increased sensitivity to insertions/deletions that DNA-duplex-based assays, such as the heteroduplex mobility assay (HMA), exhibit (Delwart et al., 2002, 1994, 1993; Sagar et al., 2004). Most of all, none of the diversity-based assays described thus far are able to distinguish between single-virus long-term infections and multiple-virus recent infections.

The presence of infections seeded by multiple founder viruses provides a significant challenge for diversity-based assays. Though the majority of infections are thought to be founded by a single virus, a non-negligible proportion are founded by two or more virions. This proportion appears to be related to mode of infection (Ritola et al., 2004; Templeton et al., 2009), and there may exist effect modifiers, such as the presence of inflammatory genital infections (Abrahams et al., 2009; Haaland et al., 2009). Various studies have yielded a remarkably consistent estimate of 80% of heterosexual transmissions involving single virus transmission (Abrahams et al., 2009; Grobler et al., 2004; Keele et al., 2008; Salazar-Gonzalez et al., 2008), although there may be a higher risk for male-to-female versus female-to-male transmission (Delwart et al., 2002; Long et al., 2000). Many reports from men who have sex with men have demonstrated very little multiplicity of infection

(Delwart et al., 1997; Gottlieb et al., 2004; Shankarappa et al., 1998), though this has not been consistent across all studies (Kearney et al., 2009; Ritola et al., 2004). As might be expected because of the lack of a mucosal barrier, parenteral transmission, such as through injection drug use, is associated with a much higher proportion of multiple infections (Bar et al., 2010; Templeton et al., 2009). Simple measures of diversity, such as the fraction of ambiguous nucleotides or the HRM score, cannot discriminate between single and multiple infections; as a result, they will suffer from reduced sensitivity because recent multiple infections will appear long-term by virtue of their high levels of diversity. Since the underlying proportion of multiple infections is related to mode of transmission, which may vary across samples, it is unlikely that these incidence estimates can be reliably adjusted for their imperfect sensitivity.

Overall, multiply infected individuals have high viral diversity, even at the earliest stages of infection. Interestingly, there is evidence to suggest that the individual lineages founded by each of the separate virions have very low initial diversity. This has been recognized in studies which have sequenced quasispecies from individuals recently infected with multiple viral strains (Long et al., 2000; Salazar-Gonzalez et al., 2008). These sublineages conform to a Poisson distribution when analyzed individually (Keele et al., 2008). Thus, the challenges associated with multiplicity of infection may be overcome if the viral population can be separated into related sublineages and the diversity measured within these clusters.

The most recent diversity-based assay described in the literature uses features of the distribution of pairwise Hamming distances of *env* sequences (Park et al., 2011). These sequences are obtained using single genome amplification and direct sequencing, a method which reduces Taq polymerase errors, Taq polymerase mediated template switching, and non-proportional representation of target sequences (Keele et al., 2008). The authors use the tenth percentile of the distance distribution to discriminate between recent and long-term infections, with the reasoning that the proportion of similar sequences will decline with time. Their method appears to be robust to a variety of different factors, including

viral subtype and multiplicity of infection. No studies have yet evaluated these claims using different data sets (Allam et al., 2011).

2.4 Discussion

Despite the immense effort to develop an accurate and reliable estimator for HIV-1 incidence, no method has emerged that is sufficiently sensitive, specific, and robust to differences in host and viral characteristics. Cross-sectional assays, which require the fewest assumptions and are logistically most feasible, are likely to be the best option, but they are not without their challenges. Immunoassays, such as the BED-CEIA and the Avidity Index, suffer from low specificity resulting from ART use, chronic disease, and the presence of long-term non-progressors in the population. Diversity-based assays are emerging as a new alternative to immunoassays. Because of the added complication of multiplicity of infection, simple measures of diversity are unlikely to be successful in distinguishing between recent and long-term infections. I believe that sequencing of the viral population using single genome amplification and direct sequencing will be necessary to fully characterize the diversity within an individual. Eventually, these assays will need to be evaluated for their performance in the presence of ART and chronic infection. Overall, it is unlikely that a single assay will work well enough for use in research or routine surveillance; thus, focus is now shifting to multiassay algorithms (MAAs) which incorporate different measures, such as viral characteristics, host immune response, and use of antiretrovirals (Brookmeyer et al., 2013a,b; Moyo et al., 2014). An accurate measure of viral diversity could be incorporated as a refining step at the end of such an algorithm to improve HIV-1 incidence estimation (Cousins et al., 2014).

3. Adapting entropy to measure within-host viral diversity for use in a cross-sectional HIV-1 incidence assay

Natalie Exner and Marcello Pagano

Abstract

HIV-1 incidence can be estimated using cross-sectional assays that discriminate between recent and long-term infections on the basis of host or viral characteristics. Because of the limitations of existing immunological assays, assays measuring within-host viral diversity are potentially useful because diversity generally increases with time since infection. One such assay described in the literature is the 10th percentile (Q10) of the within-host pairwise Hamming distance distribution. A standard measure of viral diversity is Shannons entropy which either quantifies the variability in the sequence patterns within the alignment or summarizes the variability at each position in the gene/region. We propose subdividing the gene/region of interest into sections of moderate length and calculating entropy as a weighted average of the entropy in each section. Such a method is hampered when an individual has been multiply infected. To overcome this shortcoming, we propose a clustering based method that separates the sample into unique sub-lineages before measuring entropy, if there is evidence of multiple infections. To evaluate our approach, we analyze envelope sequence data from a longitudinal subtype C infection cohort in Botswana comprised of 8 acute and 34 recent infections. Sequences were obtained using single genome amplification followed by direct sequencing. We demonstrate that using sections of moderate length and adjusting for multiple infections results in significantly improved discriminatory ability of the assay relative to either pre-existing entropy-based measure (using the whole alignment or each position) or the Q10 method.

3.1 Introduction

An accurate estimator of HIV-1 incidence is critical for understanding transmission dynamics, evaluating the performance of prevention programs, and identifying high-risk populations for targeted interventions (Rutherford et al., 2000). Cross-sectional surveys for incidence estimation are promising for a variety of reasons, including increased feasi-

bility and reduced selection bias as compared to direct observation of an HIV-1 negative cohort (Brookmeyer and Quinn, 1995). To accurately estimate incidence cross-sectionally, an assay must be able to distinguish between recent and long-term infections on the basis of a measured host or viral marker. Current HIV incidence assays that rely on changes in host immunological factors during early stages of infection, such as the BED-CEIA (Parekh et al., 2002) and the Avidity Index (Suligoi et al., 2003), have critical limitations, including misclassification of those with advanced infection (Karon et al., 2008; Marinda et al., 2010), misclassification of those on ART (Marinda et al., 2010), and variability in performance across subtypes (Karita et al., 2007; Parekh et al., 2011).

An emerging alternative to immunoassays are diversity-based assays that exploit the increase in viral diversity that occurs during the early stages of HIV-1 infection. Numerous studies have demonstrated that within-host viral diversity increases over the course of HIV infection (Frost et al., 2005; Kearney et al., 2009; Lee et al., 2008, 2009; Shankarappa et al., 1999); thus, within-host viral diversity is a potential predictor of time since infection. A range of diversity-based assays have been described in the recent literature, including a measure of the fraction of ambiguous nucleotide calls obtained during bulk sequencing (Kouyos et al., 2011) and a high-resolution melting (HRM) assay that measures diversity without sequencing using the melting characteristics of DNA duplexes (Towler et al., 2010). These assays are readily implementable, but one key limitation is their inability to accurately classify individuals recently infected with multiple viruses. While the initial diversity in a single infection is zero, this initial diversity can be very high when there are multiple founding viruses, thereby confounding the relationship between diversity and time since infection. It is estimated that 20% of heterosexual infections are founded by multiple viruses (Abrahams et al., 2009; Keele et al., 2008; Salazar-Gonzalez et al., 2008), with higher rates observed for parenteral infections (Bar et al., 2010; Templeton et al., 2009). Another recently described diversity-based assay uses the 10th percentile of the distribution of within-host pairwise Hamming distances obtained via single genome amplification (Park et al., 2011). The authors suggest that this method has high sensitivity

and specificity and is robust to the presence of multiple infections.

We propose an alternative measure of within-host viral diversity also using data from single genome amplification and direct sequencing. Our measure is an adaptation of a standard measure of viral diversity – normalized Shannons entropy. Entropy is a concept from information theory that measures variability in a multinomial outcome (Shannon, 1948). There are two common ways in which entropy is calculated for viral sequences. In the first approach, the multinomial outcomes are the observed viral sequence patterns in an alignment (henceforth referred to as the whole alignment approach) (Sato et al., 1998; Wang et al., 1998). Entropy (H) can be defined as:

$$H = -\left(\frac{1}{\log N}\right)\sum_{i=1}^{n} p_i \log p_i$$

where N is the total number of sequences, n is the number of distinct sequence patterns, and p_i is the proportion of sequences consisting of each distinct sequence pattern. The measure is normalized by the quantity $\log N$, which is the maximal entropy for a set of N sequences. Entropy is equal to 0 when all sequences are identical and is equal to 1 when all sequences are distinct.

The second approach entails calculating the entropy at position j, where n is the number of distinct bases (or amino acids) at each position and p_{ij} is the prevalence of base i (or amino acid i) at position j; the entropy of the alignment is then the average entropy over all positions (henceforth referred to as the position approach) (Korber et al., 1994).

We argue that it is possible to improve the ability of entropy to discriminate between recent and long-term infections by modifying how the quantity is calculated. We demonstrate that the whole alignment approach described above is not well-suited for discrimination because, by this definition, two sequences varying by only a single nucleotide have distinct patterns. As a result, this method is not sufficiently nuanced to measure the range of similarities that can occur between sequences. In addition, we demonstrate that the second version of Shannons entropy the position approach is also not well-suited for discrimination because it over-represents highly correlated positions. We propose a novel

method that subdivides the viral gene into smaller sections, calculates entropy in each individual section, and combines results across all sections. The optimal section length is a compromise between the two extreme lengths defined by the whole alignment and the position methods, and can be determined by its discriminating capabilities.

As can be anticipated, a multiply-infected individual will exhibit more variability than a singly-infected individual, both infected at the same time, and as a result, judging time since infection by looking at the diversity will be complicated by the multiplicity of the infection. To address this problem, we also propose an approach for identifying multiple infections using standard clustering techniques. We suggest that viral diversity be measured within distinct sub-lineages rather than across all sequences to reduce the confounding effect of multiple infections. This approach is motivated by evidence that the separate sub-lineages in a multiple infection evolve independently (Keele et al., 2008). Thus, the diversity within each sub-lineage may be comparable to the diversity accumulated within a single infection. We evaluate our proposed methods using data from a primary HIV-1 subtype C infection cohort in Botswana (Novitsky et al., 2009). The cohort includes eight acutely infected and thirty-four recently infected individuals followed longitudinally through 500 days post-seroconversion. We demonstrate improved discriminatory ability of our approach when comparing it to the 10th percentile (abbreviated as Q10) (Park et al., 2011) of the pairwise Hamming distance distribution.

3.2 Materials and methods

3.2.1 Sequence data

We analyze viral sequences from the Tshedimoso study, a primary HIV-1 subtype C infection cohort in Botswana comprised of eight acutely infected individuals (Fiebig stage II) and thirty-four recently infected individuals (Fiebig stage IV or V) (Novitsky et al., 2009). For acutely infected subjects, time of seroconversion was estimated as the mid-

point between the last seronegative test and the first seropositive test; for recently infected patients, time of seroconversion was estimated using Fiebig staging (Fiebig et al., 2003).

Subjects were followed longitudinally through 500 days post-seroconversion. At each time point, sequences from the envelope gene were obtained using single genome amplification followed by direct sequencing (Novitsky et al., 2009). After aligning the sequences, hypermutants were removed from the sample using Hypermut 2.0 (Rose and Korber, 2000). The total length of the alignment was 1377 nucleotides. Only samples with at least 5 sequences collected were considered for analysis. The 42 subjects were observed at a total of 197 time points with a median of 11 sequences per time point (range 5, 32).

From the aligned sample of HIV-1 viral sequences, for each subject at each time point we calculate the 10th percentile of the pairwise Hamming distance distribution (Q10).

3.2.2 Proposed method

Our methods require an aligned sample of HIV-1 viral sequences obtained cross-sectionally from a single host. Each sequence can represent a single region, multiple regions, a single gene, or multiple genes. For the sequences to be representative of the underlying viral population, they are obtained using single genome amplification followed by direct sequencing.

(1) Sectioned entropy

Given an alignment of sequences with multiple sequences per individual, we subdivide the aligned sample into smaller sections. For each individuals set of aligned sequences, we calculate the entropy separately for each section, and then we average across the sections to obtain an overall score (see Figure 3.1). By calculating diversity in this way, isolated polymorphisms do not disproportionately contribute to entropy because they are countered by the lack of diversity in other sections. The score is a weighted average with weights equal to the length of the section (after removing all gap-only sites in the sam-

ple). The length and location of these sections can be determined in a variety of ways. We select a simple division of the sample into approximately even sections of L nucleotides, testing a variety of values for L (ranging from 50 to 500), but the method can easily be adapted for more biologically motivated divisions, such as breaking up the sample into functional domains, or flexible weighting of regions.

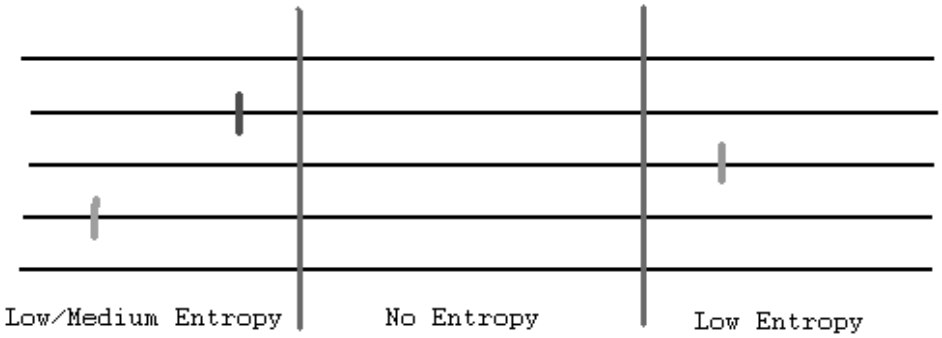

Figure 3.1: Highlighter plot example of sectioned entropy procedure. The alignment of sequences within a host is subdivided into sections of a pre-determined length. Entropy is measured within each section, and the overall entropy score is a weighted average of the section-specific entropy measures.

(2) Adjusting for multiplicity of infection

Next, we propose an algorithm to reduce confounding due to the presence of multiple infections. Multiple infections can be characterized by a high degree of clustering among the sequences within a patient. After dividing the alignment into sections, we test each individuals sequences for the presence of multiple infections using clustering methods. If present, we separate the sample into distinct sub-lineages, measure the entropy in each sub-lineage, and combine information across sub-lineages to obtain the entropy for that section (see Figure 3.2). By measuring entropy within sub-lineages rather than across sub-lineages, we reduce the risk of over-estimating diversity at early stages of infection when we would expect low variability within sub-lineages but potentially high variability across sub-lineages. Because within-host recombination can dilute the strength of clustering, clustering is assessed in each section separately to allow sequences within an individual to group with different sub-lineages across sections. The clustering procedure is

repeated for all sections, and the overall score is the average entropy across all sections.

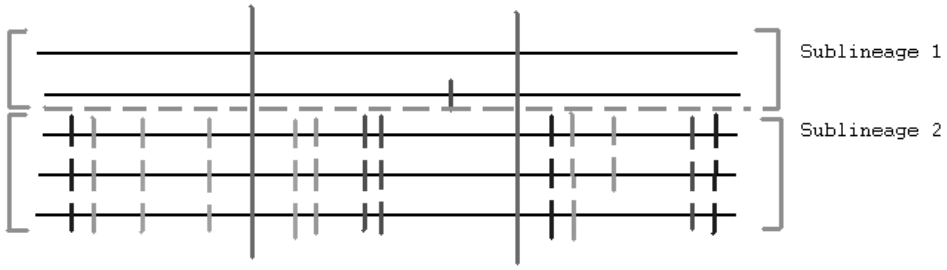

Figure 3.2: Highlighter plot example of adjustment for multiplicity of infection procedure. Within each section, clustering methods are applied to separate sequences within a host into distinct sub-lineages. Entropy is measured within each sub-lineage and combined across sub-lineages.

To test for the presence of clustering within an individual in an aligned section, we use Hamming Distance (or any other genetic distance metric) to calculate a matrix of pairwise distances for all of the sequences within a host, and then we compare the maximum observed pairwise distance to a moderately low threshold (3%) because sequences with a maximum pairwise distance below this are unlikely to be multiply infected. Among sequences with a sufficiently high maximum pairwise distance we use an automatable procedure employing a measure known as the silhouette width to determine the optimal number of clusters given the observed data (Kaufman and Rousseeuw, 1990). This procedure can return a value of 1, 2, 3 or 4 clusters, with 1 cluster indicating no significant clustering. To apply this procedure, first use the k-means algorithm on the matrix of within-host pairwise distances to cluster sequences into k = 2, 3, and 4 groups (Kaufman and Rousseeuw, 1990). For each value of k, calculate the average silhouette width. Silhouette width is a measure of the adequacy of clustering that can take values between 0 and 1, with an average silhouette width above 0.70 indicating a strong structure. Average silhouette width maximizes when the number of clusters is optimal. Silhouette width is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

where, for sequence i, $a(i)$ is the average distance to other sequences in its cluster, and $b(i)$ is the average distance to sequences in the nearest cluster, with nearest defined as

having the minimum average distance to sequence i. The average silhouette width is the average over all sequences i. Since silhouette width is not calculated for $k = 1$, the optimal number of clusters $k*$ is the value of k that yields the maximum silhouette width, conditional on that maximum being above 0.70; otherwise, $k*$ is equal to 1.

Given the calculated number of clusters $k*$ for that section, we subdivide the sequences into sub-lineages using k-means clustering. We then measure entropy in each sub-lineage, and these entropy scores are pooled across sub-lineages, weighting proportional to the number of sequences in each sub-lineage minus 1 (since diversity is trivially zero from a sample of size 1 and should not contribute to the overall diversity). Letting H_K be the entropy in the kth sub-lineage, the overall entropy in that section is:

$$H = \frac{\sum_{k=1}^{k*}(n_k - 1)H_k}{\sum_{k=1}^{k*}(n_k - 1)}$$

As before, the overall entropy score is a weighted average of the entropy values across all sections, weighted by section length.

3.2.3 Statistical analyses

We calculate each approachs discriminatory ability at 180 days post-seroconversion via the area under the curve (AUC) of the receiver operator characteristic (ROC) curve and an associated 95% confidence interval. We employ a non-parametric approach to adjust for clustering of the data by subject because subjects are observed repeatedly over time (Obuchowski, 1997). To directly compare two different approaches using the same sample of clustered data, we apply a similar method to calculate the absolute difference in AUC and a 95% confidence interval for this difference (Obuchowski, 1997).The analyses were carried out using R (R Core Team, 2012).

3.3 Results

For all subjects at all time points, we calculate entropy using the whole alignment approach, the position approach, and the section approach, dividing the alignment into sections roughly of length $L = 50$ through $L = 500$ nucleotides (in increments of 50) with and without adjusting for multiplicity of infection using the Silhouette method described.

In Figure 3.3 we plot entropy score trajectories for each of the 42 subjects against time since seroconversion using whole alignment entropy ($L = 1377$) and sectioned entropy for three additional values of L (50, 300, and 500 nucleotides) without adjusting for multiplicity of infection (plotted with LOESS curve 95% confidence band). Although there is heterogeneity in the level of diversity across individuals, the trajectories tend to increase with time for all measures. The whole alignment approach attains a maximal value even at the earliest time points for some samples. After introducing sectioning, the entropy scores decrease and are less likely to attain the maximal value. As a result, we observe a more pronounced relationship between time and diversity. When the sections are very small (i.e., $L = 50$), the entropy scores tend to be lower because there are more gene regions with few or no mutations.

In Figure 3.4 we plot diversity trajectories for each of the 42 subjects using Shannons entropy for $L = 250$ with and without adjustment for multiplicity of infection against time since seroconversion. The adjustment method results in slightly decreased variability in the trajectories, as evidenced by a tightening of the 95% confidence interval band. Even after adjusting for multiplicity of infection, some early infections with high entropy scores persist.

Table 3.1 summarizes the results of the clustered ROC analyses measuring discriminatory ability at 180 days post-seroconversion using the whole alignment approach, the position approach, and the section approach, dividing the alignment into sections of length $L = 50$ through $L = 500$ nucleotides (in increments of 50) with and without adjusting for

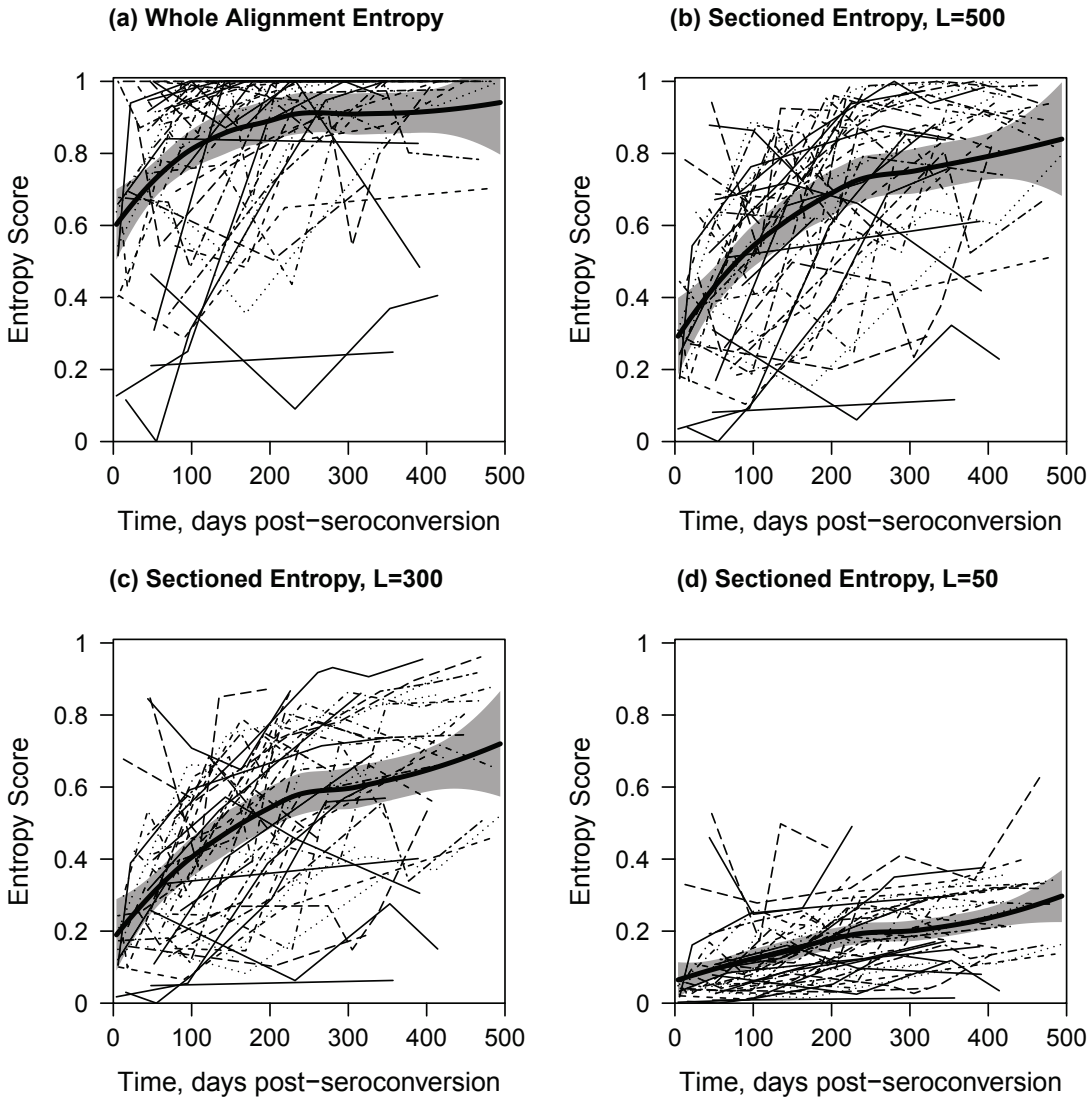

Figure 3.3: Entropy score trajectories and LOESS line for the env gene of 42 acutely and recently HIV-infected subjects. (a) Whole alignment approach. (b) Sectioning procedure is applied with $L = 500$ without adjustment for multiplicity of infection. (c) Sectioning procedure is applied with $L = 250$ without adjustment for multiplicity of infection. (d) Sectioning procedure is applied with $L = 50$ without adjustment for multiplicity of infection.

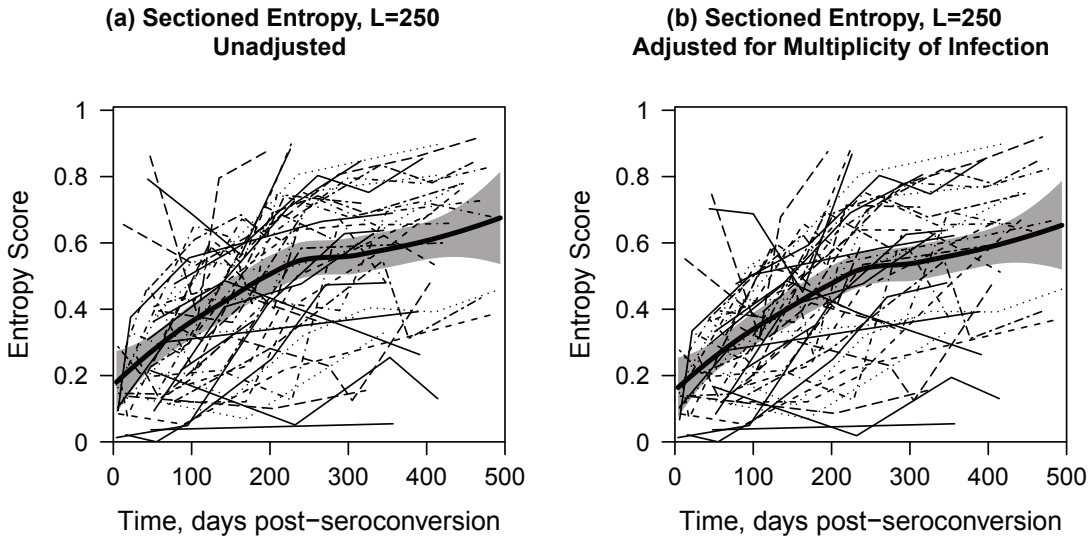

Figure 3.4: Entropy score trajectories and LOESS line for the env gene of 42 acutely and recently HIV-infected subjects. (a) Sectioning procedure is applied with $L = 250$ without adjustment for multiplicity of infection. (b) Sectioning procedure is applied with $L = 250$ with adjustment for multiplicity of infection.

multiplicity of infection using the Silhouette method described. Note that the Silhouette method is not applied to the position approach because clustering cannot be assessed in a region of length $L = 1$. Among the methods considered, the minimal AUC is 68.3% for the whole alignment approach adjusting for multiplicity of infection, and the maximal AUC is 79.3% for the sectioned entropy approach with $L = 250$ adjusting for multiplicity of infection. The AUCs for the whole alignment approach are lower than any AUC obtained via the sectioned approach. With one exception ($L = 50$ with adjustment), all combinations of the sectioned approach have a significantly higher AUC than the whole alignment approach. The AUCs for the position approach are lower than any AUC obtained via the section approach. This difference is statistically significant for $L = 150$, 200, and 250 (without adjustment) and $L = 250$ (with adjustment). Note that statistical significance is determined using the matched AUC procedure described previously; thus, certain approaches do not achieve statistical significance despite having higher AUC values. Furthermore, the statistical procedure incorporates the observed covariance between the approaches. Because the approaches use the same underlying data, the covariance can

be very high. Thus, a significant difference can be detected between methods even when there is a lot of overlap in the confidence intervals.

Table 3.1: Estimated area under the curve for discriminating between infections ¡180 and 180 days post-seroconversion on the basis of within-host viral diversity in env. Diversity is measured using whole alignment entropy, sectioned entropy, or position entropy, either adjusted or unadjusted for multiplicity of infection.

	Not Adjusted for Multiplicity of Infection		**Adjusted for Multiplicity of Infection**	
	AUC	**95% CI**	**AUC**	**95% CI**
Whole Alignment (L = 1377)	68.8%	(61.5%, 76.0%)	68.3%	(61.0%, 75.6%)
Sectioned (L = 500)	77.8%*	(70.8%, 84.8%)	78.3%*	(71.5%, 85.2%)
Sectioned (L = 450)	78.0%*	(71.0%, 85.0%)	78.9%*	(71.9%, 85.8%)
Sectioned (L = 400)	77.8%*	(71.0%, 84.7%)	78.6%*	(71.8%, 85.5%)
Sectioned (L = 350)	77.8%*	(70.9%, 84.7%)	78.3%*	(71.7%, 85.0%)
Sectioned (L = 300)	76.8%*	(69.8%, 83.8%)	77.4%*	(70.3%, 84.4%)
Sectioned (L = 250)	78.4%*†	(71.3%, 85.6%)	79.3%*†	(72.3%, 86.2%)
Sectioned (L = 200)	77.9%*†	(70.7%, 84.9%)	79.0%*	(72.0%, 86.0%)
Sectioned (L = 150)	77.9%*†	(70.5%, 85.2%)	78.8%*	(71.9%, 85.6%)
Sectioned (L = 100)	76.7%*	(69.4%, 83.9%)	78.2%*	(71.0%, 85.4%)
Sectioned (L = 50)	76.2%*	(68.6%, 83.8%)	74.7%	(66.9%, 82.4%)
Position (L = 1)	73.7%	(66.1%, 81.3%)	n/a	n/a

* indicates that the AUC is significantly higher than that of the unadjusted whole alignment approach.
† indicates that the AUC is significantly higher than that of the unadjusted position approach.

In Figure 3.5 we plot diversity trajectories for each of the 42 subjects using the Hamming Distance Q10 approach. Q10 tends to increase with time since seroconversion in this population. We observe some individuals with high measurements at early time points. These patients are almost exclusively patients with documented multiple infections (Novitsky et al., 2011). The sample includes one very high outlier at 469 days post-

seroconversion for a patient with a multiple infection. There are also patients with low values of Q10 persisting beyond 300 days post-seroconversion. The AUC for the Q10 method in this sample is 74.5% with 95% confidence interval (67.2%, 81.9%). The whole alignment approaches (unadjusted and adjusted) have a significantly lower AUC than the Q10 method. Nearly all combinations of the sectioned approach have a significantly higher AUC than the Q10 method (unadjusted $L = 200$, 250, 350, 400, 450, and 500, and adjusted $L = 100$, 150, 200, 250, 350, 400, 450 and 500). The maximal difference is with the sectioned entropy with $L = 250$ and adjusting for multiplicity of infection; here the absolute difference is 4.7%, with 95% confidence interval (1.9%, 7.5%).

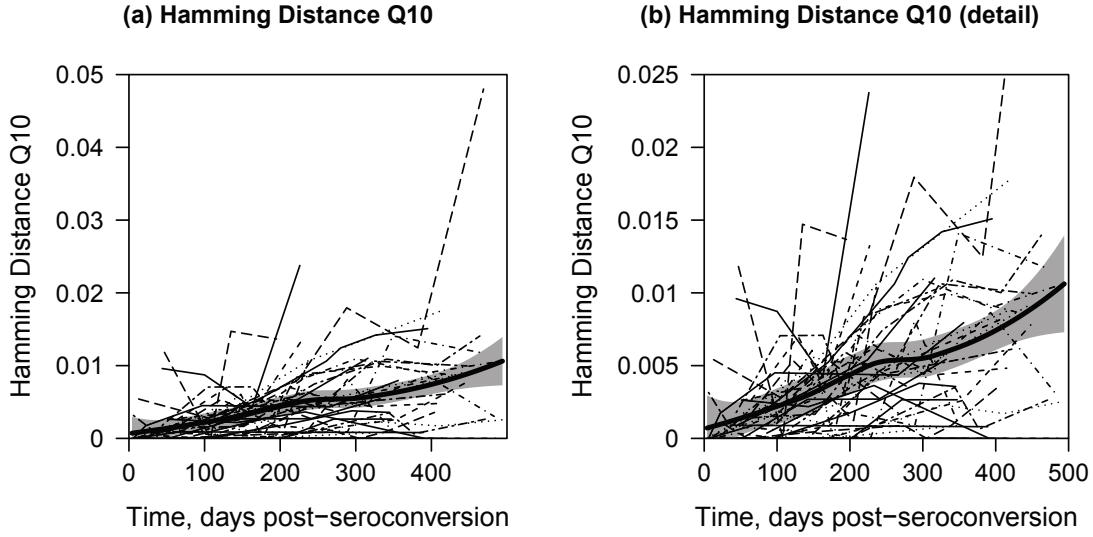

Figure 3.5: Q10 trajectories and LOESS line for the env gene of 42 acutely and recently HIV-infected subjects. (a) All trajectories. (b) Detail.

3.4 Discussion

We propose a new measure of within-host HIV-1 viral diversity for use in cross-sectional incidence estimation. To generate this measure, we describe two simple adjustments to Shannons entropy with the goal of improving our ability to discriminate between infections before and after 180 days post-seroconversion. For the first proposed adjustment,

we divide the alignment into sections of length L (recommended $L = 250$) and calculate overall entropy as an average of the section-specific entropy values. For the second adjustment, we use clustering methods to separate samples into distinct viral sub-lineages before measuring entropy. Both of these adjustments improve the discriminatory ability of entropy as evidenced by changes in AUC. We evaluate an existing diversity-based measure using the 10th percentile (Q10) of the within-subject pairwise Hamming distance distribution. We found that most combinations of sectioned entropy had significantly improved discriminatory ability over Q10 in this cohort of acutely and recently infected subtype C patients from Botswana.

The results of our research provide insight into the relationship between diversity and the size of the gene/region of interest. Traditionally, entropy is either measured as the variability in the patterns in the whole alignment, or entropy can be calculated as the average entropy over all positions in the gene/region. Both of these approaches represent extreme values for L, the length of the region in which we calculate entropy; in the former, L is equal to the total length of the gene/region; in the latter, L is equal to 1. Our research suggests that the optimal L is a compromise between these two extremes. If L is too large, we fail to capture similarities between sequences, and a sample can reach maximal entropy even if it is very homogeneous. If L is too small, we over-represent highly correlated polymorphisms, leading to undesirable behavior of the measure and poorer prediction. If we select a larger L, related mutations that fall in the same section do not contribute disproportionately to overall diversity.

There are limitations to this method. First, there is heterogeneity in the accumulation of diversity across individuals. As a result, this method is unlikely to be useful for predicting time since infection in individuals (i.e. clinical use), but it has applicability at the population-level, such as in the estimation of HIV-1 incidence. Second, this method requires single genome amplification and direct sequencing for each sample, tools which are expensive and time-intensive. We believe that (1) the financial and labor costs will decrease as technology improves, and (2) that the lack of reliable alternatives at any cost,

despite the importance of the problem, suggests that this research is worth pursuing. In addition, we propose that this assay is best used as part of a multi-assay algorithm (MAA) in which patients are screened using less expensive cross-sectional assays, such as viral load testing, BED-CEIA, or antibody avidity (Brookmeyer et al., 2013a); thus, genotyping would only be necessary for the small fraction of individuals who test as recent on all screening assays, and the measurement of diversity would be a refining step in the algorithm.

An additional limitation is that, while the adjustment for multiplicity of infection does slightly improve measurement of some multiple infections, some early infections have persistently high diversity levels even after adjustment. These cases tend to be individuals infected with multiple related viruses (Novitsky et al., 2011) that do not clearly cluster and are very difficult to distinguish from long-term single infections without additional information. Furthermore, inter-lineage recombination does occur in multiple infections. The motivation for evaluating clustering in each section rather than evaluating clustering in the entire alignment is the presence of recombination, which results in recombinant sequences clustering with different sub-lineages in different regions. Dividing the sample into sections before measuring clustering reduces the impact of recombination, but recombination breakpoints which do not fall exactly in line with section breakpoints can weaken the approach. We tested incorporating information on estimated recombination breakpoints as identified by the program RDP3 (Martin et al., 2010), but this adds an additional layer of complexity without improving overall performance (data not shown). There may be a better way to incorporate this information.

Our procedure can be generalized for other methods of measuring diversity. For example, the discriminatory ability of Q10 may improve if the gene/region is first divided into smaller sections and/or separated into distinct sub-lineages using clustering methods. This question merits further investigation.

4. WHO HIV drug resistance surveillance consulting project report

Natalie Exner and Marcello Pagano

Abstract

In this report, we describe our large-scale consulting project for the World Health Organization (WHO), focusing on the statistical challenges that arose and solutions identified. The WHO is currently in the process of redesigning their guidance for the surveillance of HIV drug resistance (HIVDR) in low- and middle-income countries. The guidance describes recommendations for monitoring two aspects of HIVDR that are of interest to country program directors: (1) resistance in individuals starting antiretroviral therapy (ART), referred to as pre-treatment drug resistance (PDR), and (2) resistance in individuals on ART for at 12 ± 3 months and ≥ 48 months, referred to as acquired drug resistance (ADR). Descriptions of our proposed approach for these two surveys is currently being prepared for publication by the WHO, and the methodology described within will then be adapted by in-country researchers into survey design protocols that will meet the particular needs of that country.

4.1 Introduction

In our work as statistical consultants designing surveillance systems for HIV drug resistance in low- and middle-income countries, we faced a variety of challenges. The first major challenge is the likely limited technical capacity of in-country researchers who will be executing these surveys. Thus, the survey design, implementation, and analysis must be exceedingly simple. The second major challenge is the extreme diversity of HIV epidemics across all low- and middle-income countries. Among the countries we have consulted with, one country has more than 4,000 HIV sites, while another has only 5. Some countries have generalized HIV epidemics, while others have epidemics that are highly concentrated among particular risk groups. Some countries have electronic medical records that have detailed information on all patients that can be queried at a national level, while others have paper-based systems that cannot be accessed without visiting

each site individually. Thus, it is important that the design is sufficiently flexible that it can be adapted for use in any country. The third major challenge is feasibility. The proposed designs can not be too costly or logistically complicated. Longitudinal follow-up, which was an element of the previous version of these protocols, is no longer considered feasible. For small countries, the designs cannot require sampling more patients than would be eligible during the survey period. Feasibility is critical because it can be the difference between the survey being implemented, or not. These are just some of the many challenges that arose during this consulting process.

Ultimately, our goal is to improve the ability of program managers to make informed decisions about their country's prevention and treatment programs through the use of surveillance data with sufficient precision and quality. To elevate the quality of the surveillance data, we employ principled statistical methodology and epidemiological expertise while being mindful of the constraints described above. Where possible, we opted for the simplest methodology without sacrificing rigor. We recognize that if we were to make the design more complex, we could get closer to the ideal statistical result. In practice, though, the more complex the design, the less likely that the survey will be done properly and the less likely that the survey will be done at all. We developed Excel-based tools with limited input required by the user to assist in the survey design process, and, with only few exceptions, we developed primary outcomes that can be easily analyzed in Stata or other survey-based statistical software. Our goal is to empower in-country researchers to collect and analyze their own data with limited external assistance to build capacity to make decisions about their national HIV programs.

The core of this report is divided into two sections, covering each of the survey design protocols. We focus on the statistical challenges and our proposed solutions. Some solutions are standard, while others required development of new methodology. We describe all in turn.

4.2 Pre-treatment drug resistance (PDR)

Section 4.2.1: Background

Section 4.2.2: Survey overview

Section 4.2.2.1: Sampling frame Construction

Section 4.2.2.2: Site stratification

Section 4.2.2.3: Site sampling

Section 4.2.2.4: Patient sampling

Section 4.2.3: sample size calculations

Section 4.2.3.1: Effective sample size

Section 4.2.3.2: Design effect due to clustering of the outcome by site

Section 4.2.3.3: Design effect due to imperfect information weighting

Section 4.2.3.4: Calculating the sample size

Section 4.2.3.5: Incorporating the finite population correction

Section 4.2.3.6: Sample size calculations when all sites are sampled

Section 4.2.4: Data analysis

Section 4.2.4.1: Site sampling weight

Section 4.2.4.2: Outcome 1a

Section 4.2.4.3: Outcomes 1b and 1c

Section 4.2.4.4: Outcomes 2a, 2b, and 2c

Section 4.2.4.5: Regional aggregation

4.2.1 Background

In high-income countries, physicians can identify appropriate treatment regimens for patients initiating antiretroviral therapy (ART) using results from routine genotypic HIV drug resistance testing. In low- and middle-income countries, the current cost of genotypic testing is prohibitive. Thus, a public health approach must be applied in which all patients are initiated on a single first-line regimen. Nationally representative surveillance of HIVDR in populations initiating ART is critical to inform the selection of an effective first-line regimen. HIVDR among patients initiating ART may be attributable to transmitted drug resistance, meaning that patients are infected with an already resistant viral strain, or resistance may be acquired due to previous exposure to ARV, in the context of prevention of mother to child transmission (PMTCT) programs, pre-exposure prophylaxis (PrEP), post-exposure prophylaxis (PEP), or previous disclosed or undisclosed antiretroviral therapy. Regardless of the origin of the drug resistance mutations, it is important for a country to understand both the prevalence and type of drug resistance circulating because of the important country and global implications for population-level treatment outcomes.

In the previous surveillance methodology described by the WHO in 2006, the pre-treatment resistance survey and the acquired drug resistance survey were part of a single longitudinal survey (Jordan et al., 2008). After an initial pilot testing period, the WHO recommended that 10 to 15 representative ART sites be sampled, and one third of these sites be surveyed each year on a three-year cycle. The survey conducted at each site was a longitudinal survey following patients through the first 12 months of therapy. Patients were assessed for HIVDR prior to ART initiation and at 12 months (or before the switch to second-line therapy). We discuss the baseline survey here and reserve discussion of the 12 month follow-up survey for the section on acquired drug resistance.

Between 2006 and 2010, forty surveys were performed in 12 countries using this standardized protocol (World Health Organization, 2012a, p. 29). During survey implementation, a variety of challenges emerged. First and foremost, the longitudinal nature of the survey made it logistically challenging, especially in settings with decentralized service-delivery models and in areas of concentrated or low prevalence epidemics (World Health Organization, 2012b, p. 7). Secondly, the duration of the survey meant that it took at least a year and a half from survey initiation until results were available (World Health Organization, 2012a, p. 32). The lag was too long to provide timely information to ART program managers. Another challenge of the previous survey was the lack of standardized guidance on how to sample sites in a representative fashion. A pilot study in Namibia was run in which sites were classified based on region and disease burden (high/low), and only one site was sampled per stratum. Design choices like this make surveys less efficient.

To increase survey feasibility, the WHO decided to replace the single longitudinal study with two cross-sectional surveys (World Health Organization, 2012b, p. 7); the first is a survey of patients initiating treatment to measure pre-treatment drug resistance (PDR), and second is a survey of patients on treatment for at least 12 ± 3 and ≥ 48 months to measure acquired drug resistance (ADR). They approached us for statistical guidance as they developed these protocols. In the following section, we describe the proposal for the PDR survey. The ADR survey is described in Section 4.3.

4.2.2 Survey overview

For the surveillance of pre-treatment drug resistance (PDR), we propose a two-stage clustered survey where the primary sampling units (PSUs) are sites where patients initiate ART, and the secondary sampling units (SSUs) are patients initiating treatment at these sites during the 6 month survey period. The survey duration was chosen to be six months because it is short enough to provide timely information to program managers but long enough so that small countries can enroll enough eligible patients. Sites are selected pro-

portional to some measure of size, as described in Section 4.2.2.1, using systematic sampling, as described in Section 4.2.2.3. Eligible patients initiating therapy are enrolled at each sampled site until a predetermined patient quota is achieved, as described in Section 4.2.2.4. These patients are asked about any prior exposure to antiretroviral drugs (ARVs), and specimen samples are genotyped to test for the presence of HIV drug resistance mutations. After the site-specific quota is achieved, sites continue to screen patients for presence and type of prior exposure, as described in Section 4.2.2.4.

The primary outcomes of the survey are listed below. Outcome 1 measures HIV drug resistance among different groups of patients. Outcome 2 measures the prevalence of prior exposure to ARVs.

1a. Prevalence of HIV drug resistance among all initiators, regardless of prior exposure to ARVs

1b. Prevalence of HIV drug resistance among ART initiators without prior exposure to ARVs

1c. Prevalence of HIV drug resistance among individuals initiating ART with NNRTI-based regimens without prior exposure to ARVs

2a. Proportion of all ART initiators without prior exposure to ARVs

2b. Proportion of all ART initiators with prior exposure to ARVs

2c. Proportion of all ART initiators with unknown prior exposure to ARVs

These outcomes were determined through discussions involving the WHO and partners. They were selected because of their relevance to national program managers.

4.2.2.1 Sampling frame construction

Prior to sampling sites, the country must construct their sampling frame. The sampling frame is a list of all ART sites in the country where patients initiate treatment and the relative sizes of these sites. Ideally, site size is estimated as the number of treatment initiators observed at that site during a recent time period, such as the previous 6 months. If sampling is performed proportional to the number of initiators enrolled at each site during a previous time period, we refer to this as Probability Proportional to Size, or PPS, sampling. Technically, the sampling is PPES (Probability Proportional to Estimated Size) because the number of initiators enrolled at each site is expected to vary over time (Yansaneh, 2005, p. 17), but we refer to it as PPS to distinguish it from the other option we present. If information on the number of initiators enrolled at each site during a previous time period is not available, the country can perform Probability Proportional to Proxy Size, or PPPS, sampling. In PPPS sampling, the proxy measure is some measure of site size, generally the number of patients enrolled at that site during a recent time period. This will not be exactly proportional to the number of initiators, but it is a reasonable alternative that will distinguish between large, medium, and small sites. We expect PPPS sampling to be less efficient than PPS sampling. We describe how the choice between PPS and PPPS affects the sample size calculations during our discussion of the design effect (see Section 4.2.3.3).

To improve feasibility of the survey, we provide guidance for the exclusion of sites that are either very small or difficult to access. Very small sites are sites that would initiate very few patients during the 6 month period. Difficult to access sites are sites that the country decides a priori that they would not reasonably be able to include if selected during sampling. This might include sites located in areas of political instability or in very remote geographic areas. Though it could induce bias to exclude either very small sites or difficult to access sites, it is better to encourage countries to make these assessments prior to sampling and to do so in a principled way. We suggest that sites that are excluded

should not represent more than 10% of the patient population. Thus, if more than 10% of the patient population is treated at very small sites, then at least some of these sites should be included in the sampling frame. The 10% threshold was identified as a compromise to improve feasibility while limiting bias. If the absolute difference in PDR prevalence between the excluded sites and included sites is 10% (which would be very high in this setting), the absolute bias would be no more than 1%.

4.2.2.2 Site stratification

We did not actively encourage stratification (also referred to as explicit stratification), in which separate sampling frames are constructed for each stratum. While we recognize that stratification can improve the efficiency of the survey, our decision stemmed from a variety of factors. The primary reason is the broad range of potential stratifying factors. Countries with generalized epidemics may be interested in very different factors than countries with concentrated epidemics, and so on. To provide usable statistical guidance and tools for survey design for all of the possible scenarios that may arise would place too much of a burden on the WHO. Suggesting a standard and relatively simple strategy increases the likelihood that the survey will be designed properly while also limiting dependence on external assistance. Other reasons for not encouraging stratification are the fact that it can be difficult to proportionally allocate sites to strata when there are few sites being sampled overall or when there are many strata. The extreme is the setting when one site per stratum is sampled, in which the efficiency gain from stratification is more or less lost because standard variance estimation becomes impossible. Finally, from the existing surveys, there was no evidence that pre-treatment drug resistance varied widely on any site-level factor (urban vs. rural, etc). Thus, it would be hard to justify the additional complexity without some suggestions of a gain in precision.

Nonetheless, we do provide some guidance on how to perform stratification for countries interested. This is available in one of the report annexes. We emphasize the importance of

limiting the number of stratifying variables, combining similar strata if any one stratum is too small, and sampling at least two sites per stratum. In our guidance, we describe how to perform sample size calculations by allocating the effective sample size proportionally to the strata. If the primary goal of the country is just to guarantee a certain degree of regional representation (i.e. at least one site per region), we provide a short method that can be used to assess how many sites must be sampled to guarantee at least one site per region when using systematic sampling with implicit stratification (see Section 4.2.2.3). Briefly, the method shows countries how to check that the systematic sampling interval is smaller than the size of the smallest region. If so, then each region will be sampled at least once regardless of the systematic sampling random starting point. If the sampling interval is too large, then either more sites must be sampled or small regions should be combined with other similar regions.

4.2.2.3 Site sampling

In the first stage of sampling, it is recommended that 15-40 sites are sampled via systematic sampling using probabilities proportional to estimated site size (Wolter, 2007, sect. 8.6). This will lead to a nationally representative selection of sites in the country. The exact number of sites to be sampled should be determined by the country. The number is a compromise between efficiency (it is more statistically efficient to sample more sites) and feasibility (it is more logistically complicated and expensive to sample more sites). In countries where there are 15 or fewer sites, all sites should be included in the survey. The survey is then a one-stage stratified survey of patients within sites, which is more efficient than a two-stage clustered survey. In countries where there are more than 15 sites, these countries have the option of including all sites or taking a sub-sample (the standard design described above).

Systematic sampling was selected because it is routinely used in surveys in the developing world (Family Health International, 2000, p. 38). It also has the added benefit

of allowing countries to employ implicit stratification to improve representativeness. In implicit stratification, the systematic sampling frame is ordered on some factor before sampling. We suggest ordering by geographic region and also size within geographic region. In this way, there is an increased likelihood that sites will be sampled from each of the geographic regions. One negative consequence of systematic sampling is that, in some countries, large sites may be sampled more than once. In this case, we ask that sites just sample proportionally more patients from these sites.

4.2.2.4 Patient sampling

In the sites sampled, consecutive eligible patients initiating ART on or after a pre-defined survey start date are enrolled until the predetermined sample size for each site is achieved. We assume that consecutive patients are independent and that there are no time trends during the 6 month period. All individuals initiating ART are eligible to be enrolled, irrespective of their prior ART history. Specimens are collected from enrolled patients prior to ART initiation, and these specimens are sent to the laboratory for HIV drug resistance genotyping. Each site should contribute roughly the same number of specimens to the sample. In practice, this may not occur because of laboratory failure or sites being too small to achieve the predetermined sample size. This is accounted for in the survey weights (see Section 4.2.4.2).

After the predetermined sample size is achieved, sites must continue to screen initiators for prior ARV exposure. The goal of this continued screening process is two-fold. First of all, this improves the precision of Outcomes 2a, 2b, and 2c, which measures the proportion of patients in each prior ARV exposure category. Second and more importantly, the estimated site sizes used for systematic sampling were approximations to improve the efficiency of sampling. In order to appropriately adjust the survey sampling weights, the sampled sites must report the actual number of ART initiators observed during the 6 month survey period. Thus, even if a site is able to achieve its patient enrollment quota in

one day, it must continue to screen initiators for 6 months. Because of serious push-back from collaborators and funding partners, a compromise was reached that sites can screen initiators for a minimum of 3 months and then extrapolate to determine the 6 month eligible population size. This requires the assumption that there are no changes in the rate of ART initiation over the 6 month survey period.

4.2.3 Sample size calculations

The survey sample size is powered to achieve sufficiently precise results for Outcome 1b, which is the prevalence of HIV drug resistance among initiators without prior exposure to ARVs. A confidence interval of half-width of $\pm 5\%$ is suggested as an appropriate compromise between feasibility and precision.

Below we provide a comprehensive description of our proposed sample size calculations and justifications for assumed values and methods used. Nonetheless, it is not necessary for in-country researchers to understand the methodology below to generate an appropriate survey design. To improve feasibility, Excel-based tools for sample size calculations were constructed. These tools require limited user input, with the majority of assumed values internalized in locked cells. We believe this will reduce the likelihood of miscalculations during the survey design process.

4.2.3.1 Effective sample size

To determine the necessary sample size for the survey, we start by determining the effective sample size for estimating the prevalence of HIV drug resistance among initiators sampled. The effective sample size refers to the number of patients, k_{eff}, we would need to sample to achieve a desired confidence interval half-width if we were conducting a simple random sample. The effective sample size is determined by the prevalence of the outcome and the desired width of the confidence interval. The effective sample size is then multiplied by the estimated design effect to yield the actual sample size of the sur-

vey.

Note: Because the method for calculating a confidence interval in the setting of clustered surveys uses a t distribution with degrees of freedom equal to the design degrees of freedom (Korn and Graubard, 1999, p. 62), our effective sample size is also a function of the number of sites sampled. When the design degrees of freedom are large (around 40 or greater), it is standard to assume that $z_{0.975} \approx t_{df,0.975}$ as this simplifies calculations. This is only appropriate when the design degrees of freedom are large. Since this design requires sampling of around 15-40 sites, the design degrees of freedom will be small, and it is thus inadvisable to make this simplification. The consequence of using this simplification would be an underestimation of the total sample size required to achieve a given confidence interval half-width.

To determine the effective sample size, consider the following formula for a Wald-type confidence interval. Here, $\widetilde{p}_{DR}$ refers to the assumed prevalence of HIVDR among initiators. Available evidence suggests that it is reasonable to conservatively assume an estimated prevalence of HIVDR among all treatment initiators of 10%. This figure, which is greater than the 5% generally reported in the literature, including in the 2012 WHO HIV Drug Resistance Report, is used as a conservative measure of expected HIVDR prevalence because higher levels of pretreatment HIVDR – approximating 10% – have been documented in some regions. n refers to the number of sites sampled, and df are the design degrees of freedom:

$$95\%\text{ CI} = \left(\widetilde{p}_{DR} - t_{df,0.975}\sqrt{\frac{\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{k_{eff}}},\ \widetilde{p}_{DR} + t_{df,0.975}\sqrt{\frac{\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{k_{eff}}} \right)$$

The design degrees of freedom are defined as df = (# of sites sampled) – (# strata) (Korn and Graubard, 1999, p. 62). If stratification is not used, $df = n - 1$. The half-width of this confidence interval is:

$$L = t_{df,0.975}\sqrt{\frac{\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{k_{eff}}}$$

The effective sample size can be calculated using the following formula:

$$k_{eff} = \frac{t^2_{df,0.975}\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{L^2}$$

k_{eff} should be rounded up to the nearest integer.

The effective sample size must be inflated by the design effect to determine the actual sample size The elements of the study design that contribute to the design effect are (1) clustering of the outcome by site ($DEFF_{clust}$, see Section 4.2.3.2), and (2) imperfect information from using data from a previous year or from a slightly different population ($DEFF_{info}$, see 4.2.3.3).

4.2.3.2 Design effect due to clustering of the outcome by site

It is first necessary to calculate the design effect due to clustering of the outcome. The similarity of HIVDR outcomes of initiators within sites is measured via the intracluster correlation coefficient, or ICC. If m is the number of patients sampled per site and ICC_{DR} is the estimated intracluster correlation for the HIVDR outcome, the design effect due to clustering can be estimated using the following formula:

$$DEFF_{clust} = 1 + (m-1)ICC_{DR}$$

In order to estimate the ICC, global data from WHO's HIV Drug Resistance Report 2012 were used. For each site in each country, the estimated probability of drug resistance for treatment initiators was used to calculate the ICC using an analysis of variance estimator (Ridout et al., 1999). Although ICC is defined as capturing the clustering of outcomes by

sites within the same country, sites in the data-set were collapsed across different countries.

For the outcome of pre-treatment HIVDR, the estimated ICC using data from the 2012 WHO Drug Resistance Report is $ICC_{DR,raw} = 0.005$. The observed prevalence of pre-treatment HIVDR in the global data is 4.5%. As the assumed prevalence of HIVDR among initiators is 10%, and since the ICC and prevalence are generally correlated, the ICC was adjusted to reflect the assumed prevalence (Guillford et al. 2005). To perform this adjustment, a linear model predicting natural log of ICC by the natural log of prevalence was applied. The equation is

$$ICC_{DR} = \exp\left\{0.91 \times \ln\left[\frac{\widetilde{p}_{DR}}{0.045}\right]\right\} \times ICC_{DR,raw}$$

It is important to note that there are limitations to these estimates. First of all, the ICC estimates are based on only the data available in the global report. A 95% confidence interval can be constructed for $ICC_{DR,raw}$ using Searle's method (Searle, 1971; Ukoumunne, 2002), and the resulting interval extends from 0.000598 to 0.0131; thus, the interval is very wide, reflecting the uncertainty in the estimate. Thus, as the survey is implemented, it is important that the data obtained be used to better inform the estimate of ICC for future iterations of the survey.

4.2.3.3 Design effect due to imperfect weighting information

As described in Section 4.2.2.1, countries may either use PPS or PPPS sampling. The survey is maximally efficient when the estimated site sizes are perfectly proportional to the true site sizes. Otherwise, there is an inflation in the variance. To estimate the effect of imperfect information on the design effect, we use a formula estimating the variance contribution for disproportionate weights (Kalton et al., 2005, eq. 23). The design effect can be approximated by $DEFF_{info} = 1 + cv^2(weights)$, where $cv()$ refers to the coefficient

of variation and weights are the survey weights. For PPS sampling, it is estimated that $DEFF_{info} = 1.10$. For PPPS sampling, it is estimated that $DEFF_{info} = 1.50$. This corresponds to inflating the sample size by 10% and 50%, respectively, to account for the imperfect information. These numbers were calculated from observing the differences in population sizes between treatment initiators and patients on ART at sites in an African country over a two year period. These numbers are approximations, and the true values may also be very country specific. As the survey is carried out, it is recommended that these values be re-evaluated and adjusted as necessary for future iterations of the survey.

The design effect is also influenced by other sources of variability. For example, different sites will have different levels of genotyping failure. This will induce additional variability in the weights. It is estimated that this source of design effect will be small, so it is ignored in the calculations to increase the simplicity of the design.

4.2.3.4 Calculating the sample size

The design effect for HIVDR is estimated using the following formula (Park and Lee, 2004):

$$DEFF = DEFF_{clust} \times DEFF_{info}$$

Given the calculated effective sample size (k_{eff}, see Section 4.2.3.1), intracluster correlation (ICC, see Section 4.2.3.2), and design effect due to imperfect information ($DEFF_{info}$, see Section 4.2.3.3), solve the following equation for m, the number of initiators to be sampled per site:

$$m = \frac{1 - ICC_{DR}}{\left[\frac{n}{DEFF_{info}k_{eff}} - ICC_{DR}\right]}$$

If such an m does not exist, or if the calculated value of m is too large to be practical in a particular setting, consider increasing the number of sites sampled, n. Because of the

design effect, sampling a larger number of sites will require fewer samples per site, and it will also require a smaller overall sample size

The sample size needs to be adjusted for two additional parameters: (i) Laboratory failure when genotyping. Based on data from the 2012 WHO HIV Drug Resistance Report, the expected genotyping failure rate is assumed to be 10%. Thus, we need to divide the required sample size by 0.90. (ii) Expected proportion of initiators without prior exposure to ARVs. In order to retain statistical power at the analysis stage when considering only patients without prior ARV exposure, the sample size needs to be adjusted for the expected proportion of initiators without prior ARV exposure. It is assumed that 75% of initiators will have not had prior exposure to ARVs, so we need to divide the required sample size by 0.75. This should be the last step in the sample size calculations.

$$m_{samp} = \frac{m}{0.90 \times 0.75}$$

4.2.3.5 Incorporating the finite population correction

Countries can apply the finite population correction at the analysis stage to reflect the fact that either a significant portion of sites in the sampling frame are included in the sample or that a significant portion of eligible patients within a particular site are included in the sample; the result is a reduction in the variance (see Chapter 5 for a more in-depth discussion of this topic). Currently, the standard method for performing sample size calculations in small countries that will be applying the finite population correction is to reduce the effective sample size (World Health Organization, 2009a). We demonstrate that this method can be inaccurate and lead to an overestimate of the sample size.

We show that the formula for the design effect due to clustering can be revised to incorporate the predicted effect of the finite population corrections which will be applied at the analysis stage. The design effect due to clustering in the absence of finite population

corrections is $DEFF_{clust} = 1 + (m-1)ICC_{VLS}$ where m is the number of patients sampled per site and ICC is the intracluster correlation. For a country with N total sites in the sampling frame and an average of $\overline{M}$ eligible patients per site, it can be shown that the design effect due to clustering can be approximated by Chapter 5:

$$DEFF_{clust} \approx \left(1 - \frac{m}{\overline{M}}\right) + \left[\left(1 - \frac{n}{N}\right) m - \left(1 - \frac{m}{\overline{M}}\right)\right] ICC_{DR}$$

The average number of eligible patients per site can be estimated as the total number of eligible patients (estimated from available data) divided by the total number of sites in the sampling frame ($\overline{M} = M/N$ where $M = \sum_{i=1}^{N} M_i$). It can be shown that the necessary number of patients per site to be sampled per site to achieve a desired precision is (Chapter 5):

$$m = \frac{1 - ICC_{DR}}{\frac{n}{DEFF_{info}k_{eff}} - ICC_{DR}\left(1 - \frac{n}{N}\right) + \frac{N}{M}\left(1 - ICC_{DR}\right)}$$

The sample size must then be adjusted for expected genotyping failure and the expected proportion of initiators without prior exposure to ARVs.

4.2.3.6 Sample size calculations when all sites are sampled

If all sites in the sampling frame will be included in the survey, the following modifications can be made to the sample size calculations (using notation previously described). Briefly, the survey effective sample size is calculated, and this effective sample size is multiplied by a design effect due to imperfect information, the expected laboratory failure, and the expected proportion of initiators without prior ARVs exposure. It is not necessary to multiply the calculations by a design effect due to clustering because all sites in the sampling frame are included in the survey. The effective sample size necessary to achieve a confidence interval of half-width L is:

$$k_{eff} = \frac{3.84\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{L^2}$$

If the finite population correction is incorporated into the calculations (where M is the total eligible population size in the country), then the effective sample size can be calculated using the following equation (Lohr, 2010, eq. 2.25):

$$k_{eff} = \frac{M \times 3.84\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}{L^2 \times M + 3.84\widetilde{p}_{DR}(1-\widetilde{p}_{DR})}$$

Because information on patient enrollment from a prior time period will be used to allocate the sample, it is recommended that the sample size be inflated slightly to account for imperfect information; this is equivalent to adjusting for a design effect for disproportionate weighting (see 4.2.3.3). Next, the sample size should be inflated by the amount of expected laboratory success rate (90%) and the expected proportion of initiators without prior ARV exposure (75%). Thus, the actual sample size for the PPS-equivalent design is:

$$k_{act} = \frac{k_{eff} \times DEFF_{info}}{0.90 \times 0.75}$$

The actual sample size is then allocated to the sites proportional to the number of eligible patients expected to be observed during the survey period. For each site, the sample size of that site is equal to the total sample size, k_{act}, times expected patient accrual at that site divided by the expected patient accrual for all sites included in the survey. For example, if 25% of patients in a country attend a particular site, 25% of the sample size should be allocated to that site. The per site sample sizes are rounded to the nearest whole number.

4.2.4 Data analysis

Data analysis is conducted using a design-based framework. We calculate each of the outcomes as a ratio, where the denominator is an estimate of the number of eligible patients

in the country during the survey period, and the numerator is an estimate of the number of such patients with the outcome of interest.

Directions for the data analysis are provided for the Stata SVY package in Stata (StataCorp, 2013). Support includes detailed instructions on how to enter the data into a spreadsheet, step-by-step directions for reading the data into Stata, and step-by-step directions for analyzing the data using drop-down menus with limited use of the command line. We also provide a sample data set and a worked out example.

Even if Stata is not used to conduct the analysis, the Stata SVY manual section on Variance Estimation contains all necessary formulae for calculating the prevalence, variance, and 95% confidence interval of each outcome (StataCorp, 2013).

4.2.4.1 Site sampling weight

Once an appropriate design is identified, sites will be sampled using either PPS or PPPS systematic sampling. In PPS, site size is estimated using prior data on the number of initiators by site. In PPPS, site size is estimated using prior data on the number of patients on ART by site. For site i, the estimated site size in the sampling frame (from either PPS or PPPS) is denoted as $\widetilde{M}_i$. If the predetermined number of sites to be selected is $n^\star$ (note that this may be different from the number of unique sites sampled, n, because sites may be sampled twice), the probability that a site is selected is equal to $n^\star\widetilde{M}_i$ divided by the total size of all sites in the sampling frame, $\widetilde{M} = \sum_{j=1}^{N} \widetilde{M}_j$. Thus, the site sampling weight is equal to the following, where $SI = \widetilde{M}/n^\star$ is the sampling interval from systematic sampling:

$$w_{site,i} = \frac{\widetilde{M}}{n^\star \widetilde{M}_i} = \frac{SI}{\widetilde{M}_i}$$

If all sites are included in the survey, the site sampling weight is equal to 1 for all sites. If a stratified survey is conducted, site weights should be constructed separately for each

sampling frame.

As described in Section 4.2.2.4, M_i is a count of the number of eligible patients attending site i observed during the 6 month survey period.

4.2.4.2 Outcome 1a

Outcome 1a is the overall prevalence of HIVDR among all ART initiators, regardless of prior ARV exposure. The site sampling weight is defined in Section 4.2.4.1. The patient sampling weight for all initiators in site i is defined as M_i divided by the number of initiators with genotyped data available from that site, m_i. The overall weight is the product of the site and patient sampling weights:

$$w_{1i} = w_{site,i} \times \frac{M_i}{m_i}$$

For a setting without stratification, the prevalence, $\widehat{p}_{1a}$, is estimated using a ratio, letting $\widehat{t}_i$ indicate the number of initiators observed with HIVDR at site i:

$$\widehat{p}_{1a} = \frac{\sum_{i=1}^{n} w_{1i}\widehat{t}_i}{\sum_{i=1}^{n} w_{1i}m_i} = \frac{\widehat{T}}{\widehat{M}}$$

The denominator of the ratio, $\widehat{M}$, is an estimate of the total number of individuals initiating ART during the 6 month survey period in the country. (If sites are excluded from the sampling frame, it is technically an estimate of the number of eligible individuals in sites in the sampling frame.) The numerator, $\widehat{T}$, is an estimate of the total number of these ART initiators with any HIVDR mutations.

The variance is calculated using Taylor series linearization. Briefly, the variance of the ratio is expressed as a linear combination of the variance of the numerator, the variance of the denominator, and the covariance of the two (Lohr, 2010, sect. 9.1). The variance of the numerator total is (StataCorp, 2013, sect. variance estimation, eq. 2):

$$\widehat{var}(\widehat{T}) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \sum_{i=1}^{n} \left(w_{1i}\widehat{t}_i - \frac{1}{n} \sum_{j=1}^{n} w_{1j}\widehat{t}_j \right)^2 + \frac{n}{N} \sum_{i=1}^{n} \left(1 - \frac{m_i}{M_i}\right) \frac{m_i^2}{m_i - 1} \frac{\widehat{t}_i}{m_i} \left(1 - \frac{\widehat{t}_i}{m_i}\right)$$

The variance of the denominator total and the covariance term follow similarly, and they are described in the Stata documentation. The formulae are also generalized for surveys with stratification. The variance of $\widehat{p}_{1a}$ is the following (StataCorp, 2013, p. 187):

$$\widehat{var}(\widehat{p}_{1a}) = \frac{1}{\widehat{M}^2} \left\{ \widehat{var}(\widehat{T}) - 2\frac{\widehat{T}}{\widehat{M}} \widehat{cov}(\widehat{T}, \widehat{M}) + \frac{\widehat{T}^2}{\widehat{M}^2} \widehat{var}(\widehat{M}) \right\}$$

A 95% confidence interval can be calculated using a standard Wald formula or by a Logit transformation (see Chapter 7 for more details). The latter is currently the default in Stata.

4.2.4.3 Outcomes 1b and 1c

Outcome 1b and Outcome 1c are subpopulation analyses of Outcome 1a. Outcome 1b is the prevalence of HIVDR among ART initiators without prior exposure to ARVs. Data analysis is conducted using the same sampling weights described for Outcome 1a. The population is restricted to patients without prior exposure to ARVs using the subpopulation command in Stata (StataCorp, 2013, sect. subpopulation estimation). Briefly, the difference between a subpopulation analysis and a conditional analysis is that the former sets survey weights equal to zero for those not in the subpopulation and the latter entirely excludes these patients entirely (West et al., 2008). This distinction is relevant when a site has, by chance, no initiators eligible for the subpopulation analysis.

4.2.4.4 Outcomes 2a, 2b, and 2c

Outcomes 2a, 2b, and 2c are the prevalence of no prior exposure, yes prior exposure, and unknown prior exposure to ARVs, respectively, among all ART initiators. The site

sampling weight is defined in Section 4.2.4.1. The patient sampling weight for site i is defined as M_i divided by the number of initiators with recorded exposure status available from that site. The overall weight is the product of the site and patient sampling weights. The prevalence of each category is estimated using a ratio.

4.2.4.5 Regional aggregation

The WHO desired a framework for aggregating data across countries in a similar region to increase precision. For example, for the global report, the WHO may aggregate data on the prevalence of HIVDR among patients from Latin American countries with prior exposure to prevention of mother to child transmission (PMTCT) drugs. Aggregating data requires the assumption that the survey designs are comparable; specifically, the surveys should be conducted in a relatively limited time frame in order for their aggregation to be defensible because of possible time trends.

This analysis can be readily achieved by treating countries as fixed strata and analyzing the data using a combined ratio estimate (Särndal et al., 2013, eq. 7.3.13) (Wu, 1985). We provide the methodology for aggregating data across H countries (indexed $h = 1, \ldots, H$). The outcome for country h is the following:

$$\widehat{p}_h = \frac{\widehat{T}_h}{\widehat{M}_h}$$

The aggregated point estimate, $\widehat{p}$, is the following:

$$\widehat{p} = \frac{\widehat{T}}{\widehat{M}} = \frac{\sum_{h=1}^{H} \widehat{T}_h}{\sum_{h=1}^{H} \widehat{M}_h}$$

Essentially, the point estimate combines information across countries, weighting by the size of the eligible population, $\widehat{M}_h$, as estimated by their national survey. In addition, this approach does not require the eligible population sizes to be known with certainty. The variance of $\widehat{p}$ can be calculated using Taylor Series linearization (as in Section 4.2.4.2):

$$\widehat{var}(\widehat{p}) = \frac{1}{\widehat{M}^2}\left\{\widehat{var}(\widehat{T}) - 2\frac{\widehat{T}}{\widehat{M}}\widehat{cov}(\widehat{T},\widehat{M}) + \frac{\widehat{T}^2}{\widehat{M}^2}\widehat{var}(\widehat{M})\right\}$$

As we assume that the surveys are independent across countries:

$$\begin{aligned}
\widehat{var}(\widehat{T}) &= \sum_{h=1}^{H}\widehat{var}(\widehat{T}_h) \\
\widehat{var}(\widehat{M}) &= \sum_{h=1}^{H}\widehat{var}(\widehat{M}_h) \\
\widehat{cov}(\widehat{T},\widehat{M}) &= \sum_{h=1}^{H}\widehat{cov}(\widehat{T}_h,\widehat{M}_h)
\end{aligned}$$

One additional issue is data sharing. If all of the raw data is available from each country, this analysis can be easily conducted in Stata treating country as a fixed stratification variable. As it is unlikely that all countries will share their raw data, the WHO can still perform this aggregation procedure as long as they have the following elements for their outcome of interest: (1) $\widehat{T}_h$, (2) $\widehat{M}_h$, (3) $\widehat{var}(\widehat{T}_h)$, (4) $\widehat{var}(\widehat{M}_h)$, and (5) $\widehat{cov}(\widehat{T}_h,\widehat{M}_h)$. These values can be readily returned from Stata using the total commands in the SVY framework (StataCorp, 2013). Thus, this represents only a few additional commands beyond the standard analysis for each outcome of interest in the global report (see Section 4.3.4).

4.3 Acquired drug resistance (ADR)

Section 4.3.1: Background

Section 4.3.2: Survey overview

Section 4.3.2.1: Sampling frame Construction

Section 4.3.2.2: Site stratification

Section 4.3.2.3: Site sampling

Section 4.3.2.4: Patient sampling for viral load suppression/HIVDR

Section 4.3.2.5: Patient sampling for retention

Section 4.3.3: Sample size calculations

Section 4.3.3.1: Calculating the sample size for viral load suppression survey

Section 4.3.3.2: Calculating the sample size for retention survey

Section 4.3.3.3: Predicted precision of adjusted viral load suppression outcome

Section 4.3.4: Data analysis

Section 4.3.4.1: Site sampling weight

Section 4.3.4.2: Outcomes 1a, 1b, and 1c

Section 4.3.4.3: Outcome 2a

Section 4.3.4.4: Outcome 2b

Section 4.3.4.5: Outcomes 3a, 3b, and 3c

Section 4.3.4.6: Outcome 4a

Section4.3.4.7: Regional aggregation

4.3.1 Background

As described in Section 4.2.1, in an effort to improve feasibility, the WHO has opted to replace the previously recommended single longitudinal survey of patients from baseline through 12 months on therapy with two cross-sectional studies. The first cross-sectional study is a baseline survey to assess pre-treatment drug resistance (PDR) among patients

initiating ART (see Section 4.2 for comprehensive description of the PDR survey). The second cross-sectional study is a survey of patients on ART to assess acquired drug resistance (ADR). We describe the ADR survey in this section.

In the previously used longitudinal protocol, patients were assessed for HIVDR prior to ART initiation and at 12 months (or before the switch to second-line therapy). Because of the longitudinal nature of the survey, some patients transferred out, died, stopped therapy, or were lost to follow-up before the survey end date. At the 12 month time point, the primary outcomes were (1) HIV drug resistance prevention (defined as viral load $<$ 1000 copies/mL), (2) HIV drug resistance, and (3) possible drug resistance (included in this category are people lost to follow-up, individuals who stopped antiretroviral therapy, those for whom drug resistance cannot be assessed, and those with viral load greater than 1000 copies/mL 12 months after therapy initiation but no drug resistance mutations detected). These outcomes excluded patients with documented transfer to other sites and documented deaths. Deaths were excluded because it was assumed that individuals who died within 12 months of the start of treatment were unlikely to have died because of drug-resistant HIV (Jordan et al., 2008, box 4). Outcomes were calculated as raw percentages among the relevant populations, and there was no adjustment for weighting or clustering.

By using a cross-sectional design, the survey has increased feasibility, and it can be adapted to include patients on therapy for longer periods of time; this would not be possible with a longitudinal study. Nonetheless, cross-sectional studies have important limitations. A cross-sectional survey excludes patients who are no longer receiving ART at the study site and therefore cannot be observed because they have died, been lost to follow-up or have stopped treatment. This "survivor bias" can significantly impact the interpretation of the primary outcome. Without accounting for within country or country-to-country variability in retention patterns, there are many important confounding factors, making it challenging to meaningfully a) assess changes in the national estimate of observed viral load suppression (VLS, viral load $<$ 1000 copies/mL) over time, b) com-

pare these estimates against a global standard, or c) compare estimates across countries. We describe our approach adjusting for this survivor bias in Section 4.3.2.

The updated acquired drug resistance survey includes outcomes related to viral load suppression, HIV drug resistance and retention outcomes. These are important outcomes for the monitoring of acquired HIV drug resistance. For patients achieving viral load suppression, it is assumed that, because the treatment regimen is successful, there is no "effective" drug resistance. If a country observes suboptimal levels of virological suppression, they may initiate additional investigations to identify the source of these failures. In addition, information on retention is collected to improve the epidemiologic utility of the viral load suppression outcome. Among patients with virological failure, the country is also interested in the proportion of patients failing first-line ART without evidence of drug resistance mutations; patients failing therapy without evidence of resistance mutations would benefit from programmatic measures aimed at improving adherence, whereas a high proportion of patients failing with drug resistance mutations might suggest a need for a change in first- or second-line treatment regimens (World Health Organization, 2012b).

4.3.2 Survey overview

The acquired drug resistance (ADR) survey protocol has significant overlap with the (pretreatment drug resistance) PDR survey protocol, and we refer readers to equivalent sections in the PDR portion of this document to reduce redundancy. For ADR surveillance, we propose a two-stage clustered survey where the primary sampling units (PSUs) are sites where adult patients receive ART, and the secondary sampling units (SSUs) are eligible patients receiving treatment at these sites during the 6 month survey period. Similar to the PDR survey protocol, sites are selected proportional to some measure of size using systematic sampling (see Sections 4.2.2, 4.3.2.1 and 4.3.2.3). Patient eligibility is determined by duration on therapy. The WHO has identified an early and a late time point

for the ADR survey. The early time point targets adults who have been on ART for 12 (± 3) months during the survey period, and the late time point targets adults who have been on ART for at least 48 months. These time points were selected based on clinical relevance, consistency with a the previous survey and a preexisting 12 month retention indicator, and feasibility; it can be difficult to enroll patients if the time windows are too narrow. Countries may choose to conduct the survey at a single time point or both time points simultaneously. This decision will depend on budget and programmatic needs.

For the measurement of viral load suppression and drug resistance outcomes, eligible patients on ART are enrolled at each sampled site until a predetermined patient quota is achieved for each time point, as described in Section 4.3.2.4. These patients are asked about any prior exposure to antiretroviral drugs (ARVs), and specimen samples are genotyped to test for the presence of HIV drug resistance mutations. After the site-specific quota is achieved, sites continue to screen patients for presence and type of prior exposure, as described in Section 4.2.2.4.

Because of the limitations of cross-sectional data described in Section 4.3.1, we have convinced the WHO of the importance of collecting data on retention. Data collection will be done by a retrospective chart review among patients at the sites sampled, and we adopt the preexisting PEPFAR/UNGASS indicator definition of 12 month retention (UNAIDS, 2011, sect 4.2). Currently, the indicator is measured via census only, and thus representative data on retention is very rarely available from low- and middle-income countries. If analyzed properly, the survey yields a nationally representative estimate of retention without requiring a census. For the early time point, we propose a method for combining information on patient VLS and retention into a new outcome that we argue has improved epidemiologic utility. The motivation, definition, and properties of this outcome are described in an additional paper (Chapter 6). Briefly, if we assume that all patients who are lost to follow-up are not virologically suppressed, we can estimate population-level VLS as the product of VLS observed among retained patients and the prevalence of retention. We do not collect data on retention for the late time point because it is open-ended,

and thus the expected prevalence of retention is not meaningful because eventually all patients are lost to follow-up or die.

The primary outcomes of the survey are listed below. Outcome 1 measures HIV drug resistance among different groups of patients. Outcome 2 measures the prevalence of prior exposure to ARVs.

1a. Prevalence of VLS (VL $<$ 1000 copies/mL) among individuals on ART

1b. Prevalence of VLS among individuals on first-line ART

1c. Prevalence of VLS among individuals on NNRTI-based first-line ART

2a. Nationally representative measure of retention at 12 months (*early time point only*)

2b. Prevalence of VLS among individuals on ART, adjusted for retention (*early time point only*)

3a. Prevalence of HIVDR among individuals on ART with VL $>$ 1000 copies/mL

3b. Prevalence of HIVDR among individuals on first-line ART with VL $>$ 1000 copies/mL

3c. Prevalence of HIVDR among individuals on NNRTI-based first-line ART with VL $>$ 1000 copies/mL

4. Prevalence of HIVDR among individuals on ART

These outcomes were determined through discussions involving the WHO and partners. They were selected because of their relevance to national program managers.

4.3.2.1 Sampling frame construction

Prior to sampling sites, the country must construct their sampling frame. The sampling frame is a list of all ART sites in the country where patients receive treatment and the

relative sizes of these sites. The relative size of the sites is equal to the number of patients on treatment at the site during a previous time period. We refer to this method of sampling as Probability Proportional to Proxy Size, or PPPS, sampling (also described in Section 4.2.2.1). Countries with sites that opened recently may have few or no patients on treatment for ≥ 48 months. If a country is only implementing the survey at the later time point, the country can choose to exclude sites that are not expected to observe patients on treatment for ≥ 48 months from the sampling frame. If a country is implementing both time points, we propose a stratification scheme to limit patient under-enrollment (see Section 4.3.2.2).

As described for the PDR survey, we provide guidance for the exclusion of sites that are either very small or difficult to access (see Section 4.2.2.1).

4.3.2.2 Site stratification

As described for the PDR survey, we do not actively encourage explicit stratification (see Section 4.2.2.2). One important exception is if a country has many recently opened ART sites and is planning to implement both survey time points. In this setting, the country risks sampling many recently opened ART sites and significantly under-enrolling patients for the late time point. To avoid this scenario, we provide statistical guidance for a stratified design in which sites are grouped into two categories indicating their ability to enroll patients on treatment for ≥ 48 months. In practice, this grouping will likely separate recently opened (new) sites and old sites. From the old sites, patients are enrolled for both the early (12 ± 3 months) time point and the late (≥ 48 months) time point (see Table 4.1). From the new sites, patients are enrolled for only the early time point. Thus, to identify a suitable strategy, the country first identifies a design such that they can enroll sufficient patients on treatment for ≥ 48 months among the old sites only. Then, the country identifies how many additional new sites they must sample to have a reasonable stratified design for the early time point. This can be achieved via an already developed Excel-base

Table 4.1: Sampling strategy for both time points

	Early Time Point 12±3 mos.	Late Time Point ≥ 48 mos.
New Sites	First Stratum	n/a
Old Sites	Second Stratum	Only Stratum

sample size calculator.

4.3.2.3 Site sampling

Site sampling is conducted in the same manner as described for PDR (see Section 4.2.2.3).

4.3.2.4 Patient sampling for viral load suppression/HIVDR

In the sites sampled, consecutive eligible patients on ART for 12 ± 3 or ≥ 48 months (depending on which time points are implemented) on or after a pre-defined survey start date are enrolled until the predetermined sample size for each time point at each site is achieved. Specimens are collected from enrolled patients, and these specimens are sent to the laboratory for viral load testing. Specimens with viral load > 1000 copies/mL are sent to the laboratory for HIV drug resistance genotyping.

After the predetermined sample size is achieved, sites must continue to screen patients to assess their eligibility for the survey. As described for the PDR survey, the primary goal is to determine the total number of eligible patients observed at that site during the survey period for proper weighting (see Section 4.2.2.4). For the early time point, this is the total number of unique patients on treatment for 12 ± 3 months observed at the site during the 6 month survey period. For the late time point, this is the total number of unique patients on treatment for ≥ 48 months observed at the site during the 6 month survey period. As for the PDR survey, we require that screening continues for a minimum of three months regardless of the time necessary to complete enrollment.

4.3.2.5 Patient sampling for retention

For the early time point only, in order to obtain a measure of retention for patients on treatment, countries perform a retrospective chart review at the sites sampled. To conduct this chart review, countries list patients at the site who will have been on therapy for exactly 12 months during the survey period. The total number of eligible records at that site must be recorded. Next, a random sample of these patients is selected for assessment of retention. This sample can be obtained via systematic sampling (e.g., every 10th record beyond a random start point) (Lohr, 2010, sect. 2.7). 12 month retention is defined as the patient being retained on ART at exactly 12 months after treatment initiation (UNAIDS, 2011, sect. 4.2). Patients who have stopped treatment, died, or been lost to follow-up are not considered retained. Patients with documented transfer to another site are excluded from the sample. The inherent assumption is that these transferred patients have the same prevalence of retention and viral load suppression after transferring care as patients who did not transfer, and excluding them from the sample will properly implement these assumptions.

4.3.3 Sample size calculations

The survey sample size is powered to achieve sufficiently precise results for Outcome 1b, which is the prevalence of VLS among individuals on first-line ART. If the early time point sample is conducted, the survey is also powered to achieve sufficiently precise results for Outcome 2a, which is the prevalence of 12-month retention among all individuals. For both outcomes, a confidence interval of half-width of $\pm 5\%$ to $\pm 6\%$ is suggested as an appropriate compromise between feasibility and precision. The survey is not powered to achieve sufficiently precise results for Outcome 3a, which is the prevalence of HIV drug resistance outcomes among patients failing therapy. While this was originally the outcome of interest, the sample sizes were prohibitively large because this outcome is only measured among patients with viral suppression failure, which is approximately

15% to 30% of the population, depending on the time point. This population also excludes all viral amplification and genotyping failures. As a result, the overall sample sizes must be very large to collect usable data on a sufficient number of patients failing therapy.

As described for the PDR survey, sample size tools are available for countries designing these surveys (see Section 4.2.3).

4.3.3.1 Calculating the sample size for viral load suppression survey

Sample size calculations for the viral load suppression portion of the survey proceed similarly to those described for the PDR survey (see Section 4.2.3).

To calculate the effective sample size for the early time point, the suggested assumed prevalence of VLS is $\widetilde{p}_{VLS} = 0.85$ with suggested precision $L = 0.05$ (see Section 4.2.3.1). For the late time point, the suggested assumed prevalence of VLS is $\widetilde{p}_{VLS} = 0.70$ with suggested precision $L = 0.06$.

In order to estimate the intracluster correlation, global data from WHO's Global HIVDR Report 2012 were used (WHO 2012a, table 9). For each site in each country, the estimated probability of viral load suppression was calculated at the 12 month time point after censoring patients with documented transfer to another site. As before, ICC is estimated using an analysis of variance estimator (see Section 4.2.3.2). Using the raw data, with observed prevalence of viral load suppression of 89% at 12 months after treatment initiation, the estimated ICC is very low ($ICC_{VLS,raw} = 0.0032$). The 95% confidence interval for this quantity is -0.001425 to 0.01339; thus, the interval is very wide, reflecting the uncertainty in the estimate. For the assumed prevalence of 85% at the 12-24 month time point, the multiplicative factor is 1.34, resulting in an estimated ICC of $ICC_{VLS,early} = 1.34 \times 0.0032 \approx 0.004$. As the assumed prevalence of viral load suppression for the 48+ month time point is 70%, the estimated ICC is $ICC_{VLS,48+} = 2.15 \times 0.0032 \approx 0.008$.

As PPPS sampling is used for this survey, the estimated design effect due to dispropor-

tionate weighting is $DEFF_{info} = 1.50$ (see Section 4.2.3.3).

To calculate the necessary sample size, the same procedure described in Section 4.2.3.4 can be used. For the ADR survey, the sample size needs to be adjusted for two additional parameters (note that this is in lieu of adjusting for genotyping failure and prevalence of prior ARV exposure as described for the PDR survey): (i) Laboratory failure when measuring viral load. For example, if we expect a 15% amplification failure rate, we need to divide the required sample size by 0.85. (ii) Expected proportion of patients sampled receiving a first-line regimen. In order to retain statistical power at the analysis stage when considering patients on first-line regimen only, the sample size needs to be adjusted for the expected proportion of patients sampled receiving a first-line regimen. For the sake of simplicity, it is assumed that 95% of patients sampled will be receiving a first-line regimen.

The same procedure for incorporating the finite population correction into sample size calculations can be used for the ADR survey (see Section 4.2.3.5).

If all sites in the sampling frame will be included in the survey, the same procedure described for the PDR survey can be applied (see Section 4.2.3.6). The necessary assumed values for each time point are provided in this section.

4.3.3.2 Calculating the sample size for retention survey

The same procedure described for the viral load suppression outcome can be used to calculate necessary sample sizes to achieve a particular confidence interval width for the estimated retention at 12 months (see Section 4.3.2.4). The following parameters should be used: estimated prevalence of retention at 12 months is assumed to be 85%, i.e. $\tilde{p}_{RET} = 0.85$. The estimated intracluster correlation coefficient from global data is $ICC_{RET,raw} = 0.0713$ with an observed prevalence of 12 month retention of 76.6%. For the assumed prevalence of 85% at the 12 month time point, the estimated ICC is $ICC_{RET} = 0.667 \times 0.0713 \approx 0.0475$. The assumed $DEFF_{info} = 1.5$ because PPPS sampling is used.

After performing the necessary calculations, the sample size should be adjusted for the expected prevalence of documented transfer, assumed to be 5%, since these patients will be censored from the calculations. Thus, the sample size should be divided 0.95. If desired, the finite population correction can be incorporated using the formulas described above. The total eligible population size is an estimate of the number of patients who initiated therapy 12 months prior to survey initiation.

4.3.3.3 Predicted precision of adjusted viral load suppression outcome

Given a particular sample size for the viral load measure, and given a particular sample size for the retention measure, the predicted variance and confidence interval width for the adjusted viral load suppression outcome (Outcome 2b) can be calculated; assumptions and derivations are provided in Chapter 6. Let m be the number of patients sampled per site for the viral load suppression measure (excluding amplification failures), let s be the number of patients per site for the retention measure (excluding documented transfers), and let M be the total number of patients who initiated treatment in the year prior to the survey initiation.

Without applying the finite population corrections, the predicted variance is the following:

$$\begin{aligned} Var(\widehat{p}_{ADJ}) \approx \frac{1}{n}\Bigg\{ & \left[ICC_{VLS} + \frac{1}{m}(1 - ICC_{VLS})\right] p_{RET}^2 p_{VLS}(1 - p_{VLS}) \\ & + \left[ICC_{RET} + \frac{1}{s}(1 - ICC_{RET})\right] p_{VLS}^2 p_{RET}(1 - p_{RET}) \\ & + \left[ICC_{VLS} + \frac{1}{m}(1 - ICC_{VLS})\right]\left[ICC_{RET} + \frac{1}{s}(1 - ICC_{RET})\right] \\ & \times p_{VLS}(1 - p_{VLS}) p_{RET}(1 - p_{RET})\Bigg\} \end{aligned}$$

If the finite population corrections are applied, the predicted variance is the following:

$$
\begin{aligned}
Var(\widehat{p}_{ADJ}) \approx \frac{1}{n}\Bigg\{ & \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{N}{M}\right)(1 - ICC_{VLS})\right] p_{RET}^2 p_{VLS}(1 - p_{VLS}) \\
& + \left[ICC_{RET} + \left(\frac{1}{s} - \frac{N}{S}\right)(1 - ICC_{RET})\right] p_{VLS}^2 p_{RET}(1 - p_{RET}) \\
& + \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{N}{M}\right)(1 - ICC_{VLS})\right] \\
& \times \left[ICC_{RET} + \left(\frac{1}{s} - \frac{N}{S}\right)(1 - ICC_{RET})\right] \\
& \times p_{VLS}(1 - p_{VLS}) p_{RET}(1 - p_{RET})\Bigg\}
\end{aligned}
$$

The predicted confidence interval half-width is then $t_{n-1,0.975}\sqrt{var(\hat{p}_{ADJ})}$. Generalizations for stratified data or settings where all sites are included are described in Chapter 6.

4.3.4 Data analysis

Data analysis for the ADR survey has the same key features as the PDR survey (see Section 4.2.4). The key difference is that Stata cannot directly calculate Outcome 2b, which is the adjusted VLS measure. Nonetheless, we have provided directions for how to conduct this data analysis in Stata with a few additional commands. A worked out example is provided in the guidance. Data analysis is conducted using a design-based framework. We calculate each of the outcomes as a ratio, where the denominator is an estimate of the number of eligible patients in the country during the survey period, and the numerator is an estimate of the number of such patients with the outcome of interest.

4.3.4.1 Site sampling weight

The calculation of the site sampling weight follows the same procedure as described for PDR (see Section 4.2.4.1)

As described in Section 4.3.2.4, M_i is a count of the number of eligible patients for the VLS/HIVDR survey attending site i observed during the 6 month survey period. As described in Section 4.3.2.5, S_i is a count of the number of eligible records for retention review at site i.

4.3.4.2 Outcomes 1a, 1b, and 1c

Outcome 1a measures population-level viral load suppression (VL<1000 copies/mL) among individuals who have been on ART for 12±3 (or $\geq$ 48) months and who have been retained in care. Outcome 1a, therefore, is not adjusted to take into account the proportion of people who no longer attend sites because they have been lost to care, have died or have stopped treatment. The site sampling weight is defined in Section 4.3.4.1. The patient sampling weight for site i is defined as M_i divided by the number of patients on treatment for 12 (or $\geq$48) months with amplified viral load data available from that site, m_i. The overall weight is the product of the site and patient sampling weights:

$$w_{1i} = w_{site,i} \times \frac{M_i}{m_i}$$

The point estimator, variance estimator, and confidence interval estimator are as described in Section 4.2.4.2.

Outcomes 1b and 1c are subpopulation analyses of Outcome 1a (see Section 4.2.4.3).

4.3.4.3 Outcome 2a

Outcome 2a measures population-level retention at 12 months (see Section 4.3.1). The site sampling weight is defined in Section 4.3.4.1. The patient sampling weight for site i is defined as S_i divided by the number of patients on treatment for 12 (or $\geq$48) months with amplified viral load data available from that site, s_i. The overall weight is the product of the site and patient sampling weights. The point estimator, variance estimator, and

confidence interval estimator are as described in Section 4.2.4.2.

4.3.4.4 Outcome 2b

Outcome 2b measures viral load suppression (VL<1000 copies/mL) at 12 months among individuals sampled, adjusted for non-retention. This estimator assumes that all patients who are not retained in care at 12 months are not achieving viral load suppression. The adjusted proportion of patients on treatment for 12 months with viral load suppression is estimated using a ratio estimator. Ratio estimators are used to construct prevalence estimates in settings where both the numerator (total number of patients on treatment for 12 months in the country with viral load suppression) and the denominator (total number of patients on treatment for 12-24 months who are still retained in care in the country) must be estimated. The following formula for a ratio estimator is used, where $\hat{p}_{VLS,i}$ is a site-specific estimate of VLS (Outcome 1a), and $\hat{p}_{RET,i}$ is a site-specific estimate of retention (Outcome 2a):

$$\hat{p}_{ADJ} = \frac{\sum_{i=1}^{n} w_{site,i} S_i \hat{p}_{VLS,i} \hat{p}_{RET,i}}{\sum_{i=1}^{n} w_{site,i} S_i}$$

The associated variance estimator is described in Chapter 6.

4.3.4.5 Outcomes 3a, 3b, and 3c

Outcome 3a measures the prevalence of HIV drug resistance among individuals sampled on ART for 12±3 (or $\geq$ 48) months with viral loads greater than 1000 copies/mL. Outcome 3a is a subpopulation analysis of the overall data because the population is restricted to those individuals without viral load suppression. The site sampling weight is defined in Section 4.3.4.1. The patient sampling weight is the same as defined for Outcome 1a (see Section 4.3.4.2). For all HIV drug resistance outcomes, we must also define a non-response weight to compensate for genotyping failure. For all individuals with observed

genotype, their non-response sampling weight is defined as the number of patients with observed viral load failure at their site divided by the number of patients with observed viral load failure and observed genotype at their site. The non-response weight assumes that genotyping failure is unrelated to the presence of HIV drug resistance mutations. For all individuals with missing genotype, their non-response sampling weight is missing. For all individuals with viral load suppression, their non-response weight is equal to 1. The overall weight is the product of the site, patient and non-response sampling weights.

For Outcome 3a, the population is restricted to patients without viral load suppression using the subpopulation command in Stata (see Section 4.2.4.3). To analyze Outcomes 3b and 3c, users can input additional subpopulation specifications using the "and" operator.

4.3.4.6 Outcome 4

Outcome 4 is the prevalence of HIV drug resistance among all individuals sampled on ART for 12$\pm$3 (or $\geq$ 48) months. Data analysis is conducted using the same sampling weights described for Outcome 3a (see Section 4.3.4.5), though the population is not restricted for Outcome 4. The point estimator, variance estimator, and confidence interval estimator are as described in Section 4.2.4.2.

4.3.4.7 Regional aggregation

Results can be aggregated as described in Section 4.2.4.5.

4.4 Discussion

In the above document we describe our proposed methodology for the surveillance of pre-treatment drug resistance (PDR) and acquired drug resistance (ADR) in low- and middle-income countries. This work is the product of a large-scale consultation project

with the World Health Organization. This methodology is currently being published by the WHO and will be implemented by countries this year. We believe that the proposed methodology is statistically rigorous while still maintaining feasibility. We have already received very positive feedback from partners that the methods are intuitive and represent a significant improvement over previous versions.

5. The use of the finite population correction in survey design for national disease surveillance

Natalie Exner, Shira Mitchell, and Marcello Pagano

Abstract

The finite population correction (fpc) is a factor that can be applied to deflate the variance in settings where a large fraction (>5%) of the eligible population is included in a survey. It is appropriate to the use the fpc when the results of the survey will not be generalized beyond the eligible population. For countries conducting national disease surveillance to inform programmatic function, the fpc can dramatically reduce the variance of survey outcomes. When designing a survey, the fpc can be ignored or a simple fpc representing the fraction of the eligible survey population sampled can be incorporated into the sample size calculations. Applying the fpc results in a decrease in the survey sample size while still achieving the desired precision. We propose a novel method for calculating the sample size for a two-stage clustered survey that predicts the magnitude of the first- and second-stage fpcs as elements of the design effect. The result is an even greater decrease in required sample size. Via a series of simulations, we demonstrate that our proposed sample size calculation method achieves the desired precision even when the required sample size is dramatically smaller than that returned by the existing methods. Our method has important implications for surveillance in resource-limited settings in which reducing the overall survey cost and increasing feasibility are especially critical to national program managers.

5.1 Introduction

When conducting national disease surveillance in a small country, the finite population correction factor (fpc) can have a dramatic effect on the estimated precision of surveillance outcomes. The fpc is used to reflect the fact that samples are taken without replacement from a finite population. The fpc is equal to one minus the fraction of the population sampled, and it ordinarily multiplies the variance estimator (Lohr, 2010, eq. 2.9). If only a small proportion of the population is sampled, the fpc will be approximately one and

can safely be ignored; otherwise, the variance will be reduced. The fpc is useful when the sample size is large relative to the population size (say, >5%). It is appropriate to use the fpc when survey results will not be generalized beyond the eligible survey population (Kish, 1965, sect. 2.3). For national disease surveillance, survey results are only used for monitoring programmatic function, and thus the fpc is appropriate in this setting. For in-country researchers with limited technological expertise, the statistical software Stata allows users to specify the fpc at each level of sampling when analyzing multi-stage surveys (StataCorp, 2013). The resulting standard error estimates are reduced accordingly, with the first stage finite population correction reducing the estimated first stage variance, and so on.

When planning to conduct disease surveillance to estimate the prevalence of an outcome by using a two-stage clustered survey, one must calculate the sample size required to achieve a certain precision. This effort is, of course, complicated by the fact that the precision is affected by the prevalence being estimated. One common approach is to assume a value for the prevalence, possibly based on historical or other relevant information, and first calculate the required sample size as if one were taking a simple random sample with replacement. This sample size, known as the effective sample size, is then multiplied by an estimate of the design effect to yield the actual survey sample size. We can thus see that the design effect measures the relative variance of a survey with a particular design, such as a two-stage clustered survey, as compared to a simple random sample (Kish, 1995).

In a small country, the actual sample sizes calculated using this approach may be excessively large, approaching or even exceeding the total number of eligible survey participants. One common solution is to calculate the effective sample size assuming that a simple random sample is conducted without replacement. The simple random sample variance is deflated by an fpc equal to one minus the effective sample size divided by the total size of the eligible population (World Health Organization, 2009a). This smaller effective sample size is then multiplied by the design effect, resulting in an overall smaller actual survey sample size. In small countries, the actual sample size calculated with the

fpc may be significantly smaller than the one without, reflecting the fact that a significant proportion of the total eligible population will be included in the survey.

In this report, we describe an alternative approach for performing sample size calculations for two-stage clustered surveys. Rather than adjusting with a single fpc equal to one minus the effective sample size divided by the total population size, we propose an approach that more accurately mirrors the ultimate survey analysis in which an fpc is applied at each stage of the design. We demonstrate how to predict the magnitude of the first and second stage fpcs and how they can then be incorporated into the sample size calculations. As might be expected, the result is an even greater decrease in the estimated sample size while still preserving the overall desired precision. We demonstrate, by simulation, that the standard approach with a single fpc tends to overestimate the sample size necessary to achieve a particular precision. This work is motivated by the development of a generalizable survey protocol for HIV drug resistance surveillance in low- and middle-income countries. In resource limited settings, cost and feasibility are major factors in survey design; thus an approach that yields smaller sample sizes while preserving overall precision is desirable and has great applicability. In Section 5.2, we describe our calculations for the prediction of the effect of the finite population correction on variance estimation. In Section 5.3, we discuss three methods for sample size calculations for two-stage clustered surveys. In Section 5.4, we describe a simulation study to compare these three methods. Finally, we discuss our conclusions in Section 5.5.

5.2 Prediction of fpc effect

5.2.1 Notation

Divide a population into N primary sampling units (PSUs), and M_i secondary sampling units (SSUs) within each PSU_i, $i \in 1, \ldots, N$ with overall population size $M = \sum_{i=1}^{N} M_i$. Let p_i indicate the PSU mean of the outcome of interest in PSU_i. Thus, the overall preva-

lence is $p = \sum_{i=1}^{N} \frac{M_i}{M} p_i$.

To estimate p, we perform a two-stage clustered survey using probability proportional to size (PPS) sampling in which larger PSUs are more likely to be sampled, and an equal number of SSUs are sampled from each PSU, expecting that the smallest PSU is larger than the required sample size per PSU. n PSUs and m SSUs per PSU are sampled without replacement. Let $\widehat{t_i}$ indicate the number of successes among the m SSUs sampled from PSU_i, and thus the observed prevalence of the outcome is $\widehat{p_i} = \widehat{t_i}/m$. Letting w be the sampling weight for each SSU selected (constant across all PSUs and SSUs), we construct a ratio estimator $\widehat{p}$ defined below (Lohr, 2010, eq. 6.33):

$$\widehat{p} = \frac{\sum_{i=1}^{n} w\widehat{t_i}}{\sum_{i=1}^{n} wm}$$

5.2.2 Infinite population setting

The variance of this estimator can be decomposed into two parts, with the first corresponding to sampling n of N total PSUs using PPS sampling, and the second corresponding to sampling m of M_i total SSUs from each selected PSU using simple random sampling. Assuming that $n \ll N$ and $m \ll M_i \ \forall i$ and that M_i and p_i are independent, the variance can be approximated as follows, where Var_{PSU} measures the variance of the PSU means (see Appendix A.1.1):

$$Var(\widehat{p}) = \left(\frac{1}{n}\right) Var_{PSU} + \left(\frac{1}{nm}\right) p(1-p)\left[1 - ICC\right]$$

The intracluster (or intraclass) correlation coefficient (ICC) provides a quantitative measure of the similarity between SSUs within PSUs (Ridout et al., 1999). The numerator of the ICC represents the between PSU variability, and the denominator represents the sum of the between PSU and within PSU variabilities (Donner and Koval, 1980, p. 1). In order to calculate variances, we assume an underlying beta-binomial model for the data; in the

first stage the PSU means are sampled from a beta distribution, and the second stage SSU outcomes are sampled from a binomial distribution with the PSU mean generated in the first stage. The ICC for the beta-binomial data is equal to the following, where Var_{PSU} is the between PSU variability (variance of the beta distribution) (Ridout et al., 1999) (see Appendix A.1.2):

$$ICC = \frac{Var_{PSU}}{p(1-p)} \tag{5.1}$$

With this definition of ICC, we then calculate the design effect, which is the ratio of the variance of the estimate of population prevalence under the survey to the variance of an estimate of population prevalence under a simple random sample. In this paper, when we refer to design effect, we are referring to what Kish calls $DEFT^2$ (Kish, 1995) because our denominator is the variance of a simple random sample with replacement. Thus, the design effect can be approximated by the following well-known equation (Appendix A.1.3):

$$DEFT^2\left(\widehat{p}\right) = 1 + ICC\left[m-1\right]$$

5.2.3 Finite population setting

When the sampling fractions are non-negligible, one can incorporate fpcs into the calculations. For the first stage of sampling, express the first-stage fpc as $(1 - n/N)$; this is consistent with Stata's method of analyzing the data. Stata assumes a simple random sample of PSUs. In reality, our sample uses PPS sampling at the first stage. Alternative finite population corrections are described elsewhere (Wolter, 2007, chap. 8, eq. 8.7.6). For simplicity and consistency with Stata, we employ the generally more conservative first stage fpc of $(1 - n/N)$ (a direct comparison of this approach and the most popular

alternative fpc is described in Appendix A.1.4), although it would be straightforward to change the first stage finite fpc in the design and analysis. For the second stage of sampling, we express the second-stage fpcs as $(1 - m/M_i)$ for each i, which is appropriate because SSUs are sampled with equal probability.

We can predict the approximate variance of this estimator including fpcs as follows (see Appendix A.1.5):

$$var\left(\widehat{p}\right) \approx \left(\frac{1}{n} - \frac{1}{N}\right) Var_{PSU} + \left(\frac{1}{nm} - \frac{1}{n\overline{M}}\right) p(1-p)\left[1 - ICC\right]$$

We can then calculate the associated design effect (see Appendix A.1.6), where $\overline{M} = M/N$ is the average PSU size:

$$DEFT^2(\widehat{p}) \approx (1 - m/\overline{M}) + ICC\left[(1 - n/N)m - (1 - m/\overline{M})\right]$$

Thus, the predicted first stage fpc $(1 - n/N)$ and predicted second stage fpc and second stage fpc $(1 - m/\overline{M})$ can be expressed as part of the design effect. Calculation of these predicted fpcs does not require additional prior information over the standard method for incorporating the finite population correction, which requires knowledge of the total population size M and total number of PSUs N. If the sampling fraction is negligible during both stages of sampling, the design effect simplifies to the familiar expression for the design effect of a clustered survey, i.e., $1 + ICC\left[m - 1\right]$.

5.3 Sample size calculations

The variance of the estimator under discussion directly impacts the sample size required to reach a desired precision, which is an important consideration at the design stage of the survey. Here we present three methods for calculating the sample size for a two-stage clustered survey. The first does not incorporate any fpc (assumes an infinite population).

The second incorporates a single population-level fpc. The third is the method we propose, and it incorporates both first- and second-stage fpcs.

As stated previously, to perform sample size calculations, we require prior information about the outcome. For estimating prevalence, we require an assumed value for the prevalence, p, and some measure of the intracluster variability of the outcome, such as the ICC. If the prevalence is unknown, we can assume that it is equal to $p = 0.50$. The ICC can be challenging to estimate in practice. If prior data are available, the ICC can be estimated using an ANOVA estimator (Ridout et al., 1999). Otherwise, we recommend surveying the literature to identify a reasonable value. To design the survey, one must also specify the desired precision; this is generally defined as a desired half-width L for a 95% confidence interval with quantile q (example: 1.96 or $t_{df}(0.975)$ where df refers to the design degrees of freedom for the survey (Korn and Graubard, 1999, p. 62)). This information is combined to calculate the effective sample size, k_{eff}. As mentioned previously, the effective sample size is then multiplied by the design effect to yield the actual sample size of the survey, k_{act}.

5.3.1 Method 1: No finite population correction

For the first sample size calculation method (denoted with a subscript 1), if no fpc is used, we can determine an expression for m_1, the number of SSUs required per PSU when sampling n PSUs (see Appendix A.1.7):

$$m_1 = \frac{q^2 p(1-p)\left[1 - ICC\right]}{L^2 n - q^2 p(1-p) ICC}$$

For this and all other methods, the per PSU sample size m_1 should be rounded up to the nearest whole number. Note that the quantile q used for the 95% confidence interval is left general. Because the method for calculating a confidence interval in the setting of clustered surveys uses a t-distribution with degrees of freedom equal to the design

degrees of freedom (Korn and Graubard, 1999, p. 62), our effective sample size is also a function of the number of PSUs sampled. When the design degrees of freedom are large (around 40 or greater), it is standard to assume that $z_{0.975} \approx t_{df,0.975}$ as this simplifies calculations. When sampling only a few PSUs, the design degrees of freedom will be small, and it is thus inadvisable to make this simplification. The consequence of using this simplification would be an underestimation of the total sample size required to achieve a given confidence interval half-width.

5.3.2 Method 2: Finite population correction in effective sample size

In the second method (denoted with a subscript 2), a finite population correction is applied while solving for the effective sample size. This method is used frequently in practice when conducting surveillance in low- and middle-income countries (Yansaneh, 2005, p. 26)(World Health Organization, 2009a, p. 29). We can determine an expression for m_2, the number of SSUs required per PSU when sampling n PSUs (see Appendix A.1.7):

$$m_2 = \frac{q^2 p(1-p) M \left[1 - ICC\right]}{n \left[L^2 M + q^2 p(1-p)\right] - q^2 p(1-p) M \left[ICC\right]}$$

5.3.3 Method 3: Finite population corrections at each stage of sampling

In the third method (denoted with a subscript 3), we incorporate first- and second-stage finite population correction factors into the design effect estimate. Using this method, we can determine an expression for m_3, the number of SSUs required per PSU when sampling n PSUs (see Appendix A.1.7):

$$m_3 = \frac{q^2 p(1-p) \overline{M} \left[1 - ICC\right]}{L^2 n \overline{M} + q^2 p(1-p) - q^2 p(1-p) \overline{M} \left[ICC\right] \left[\left(1 - \frac{n}{N}\right) + \frac{1}{\overline{M}}\right]}$$

5.4 Simulations

We directly compare these three sample size calculation methods using a simulation study, varying the size and distribution of the population, the intracluster correlation coefficient, and the desired precision. We focus on the selection of HIV clinics (which are our PSUs) within a country for the purpose of estimating national HIV drug resistance prevalence. For each simulation, we generate N PSUs with M_i SSUs in each PSU_i. We then simulated the outcomes using a beta-binomial distribution, for which PSU_i has prevalence p_i drawn from a $Beta(\alpha,\ \beta)$ distribution, and each of the secondary sampling units (SSUs) in that PSU are drawn from a Bernoulli distribution with success probability p_i. To simulate data with a particular overall prevalence p and intracluster correlation ICC, the parameters from the Beta distribution must equal the following (Ridout et al., 1999, p. 138):

$$\begin{aligned} \alpha &= \left[\frac{1-ICC}{ICC}\right] p \\ \beta &= \left[\frac{1-ICC}{ICC}\right] (1-p) \end{aligned}$$

To simulate two stage cluster sampling, we randomly sample n PSUs using probability proportional to size (PPS) sampling without replacement; then, we randomly sample m SSUs from each selected PSU using simple random sampling without replacement. If m is larger than the number of SSUs in the selected PSU, sampling stops after all available SSUs are included. For each simulated cluster sample, we calculate the estimated prevalence, $\widehat{p}$, which is the mean of the observed data since the overall design is PPS and the data is self-weighting. If any clinics under-enroll, making the design no longer epsem, m_i may vary across clinics and the sampling weights w_i for each individual will be equal within clinics but not across clinics. The slightly more general formula below can be used to estimate the prevalence:

$$\widehat{p} = \frac{\sum_{i=1}^{n} w_i \widehat{t}_i}{\sum_{i=1}^{n} w_i m_i}$$

Following the Stata SVY documentation (StataCorp, 2013, "variance estimation"), an estimator of the variance can be written as:

$$\begin{aligned}\widehat{var}(\hat{p}) &= \left(1 - \frac{n}{N}\right) \frac{1}{\hat{M}^2} \frac{n}{n-1} \sum_{i=1}^{n} w_i^2 \left(\hat{t}_i - \hat{p} m_i\right)^2 \\ &\quad + \frac{1}{\hat{M}^2} \frac{n}{N} \sum_{i=1}^{n} \left(1 - \frac{m_i}{M_i}\right) \frac{m_i}{m_i - 1} w_i^2 \left\{m_i \hat{p}_i (1 - \hat{p}_i)\right\}\end{aligned}$$

This variance has associated 95% Wald confidence interval:

$$\left[\widehat{p} - t_{n-1,0.975} \sqrt{\widehat{var}(\widehat{p})},\ \widehat{p} + t_{n-1,0.975} \sqrt{\widehat{var}(\widehat{p})}\right]$$

We ran 25,000 iterations for each of six scenarios. The confidence interval (CI) width reported is the average confidence interval half-width over the 25,000 iterations.

Scenario 1 (*moderate ICC, small number of large clinics*): $ICC = 0.01$, $p = 0.80$, desired CI width is $\pm 5\%$ ($L = 0.05$). $N = 30$ clinics, each of size $M_i = 100$. $M = 3000$.

Table 5.1: Average CI width for sampling $n = 15$ and $n = 20$ clinics in Scenario 1

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 15$	$m_1 = 25$ $k_{act1} = 375$	± 0.0403	$m_2 = 22$ $k_{act2} = 330$	± 0.0433	$m_3 = 18$ $k_{act3} = 270$	± 0.0485
$n = 20$	$m_1 = 17$ $k_{act1} = 340$	± 0.0427	$m_2 = 15$ $k_{act2} = 300$	± 0.0458	$m_3 = 13$ $k_{act3} = 260$	± 0.0496

The per clinic sample size is smaller for Method 3 (our proposed method) than for Methods 1 (no fpc) and 2 (simple fpc), thus representing a great savings in cost and time. This saving is quite sizable when only $n = 15$ clinics are sampled, requiring 60 fewer samples overall than the simple fpc method The sample sizes calculated from Method 3 result in confidence intervals at approximately the desired $\pm 5\%$ precision. Because they are larger, the sample sizes calculated from Methods 1 and 2 yield confidence intervals narrower than the prescribed $\pm 5\%$.

Scenario 2 (*moderate ICC, large number of small clinics*): $ICC = 0.01$, $p = 0.80$, desired CI width is $\pm 5\%$ ($L = 0.05$). $N = 100$ clinics, each of size $M_i = 30$. $M = 3000$.

Table 5.2: Average CI width for sampling $n = 15$ clinics in Scenario 2

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 15$	$m_1 = 25$ $k_{act1} = 375$	± 0.0274	$m_2 = 22$ $k_{act2} = 330$	± 0.0320	$m_3 = 14$ $k_{act3} = 210$	± 0.0479

Again, the per clinic sample size is smaller for Method 3 ($m_3 = 14$ for $n = 15$) than for Methods 1 and 2 ($m_1 = 25$ and $m_2 = 22$, respectively for $n = 15$). Note that Methods 1 and 2 result in the same sample sizes for Scenarios 1 and 2 as they have the same ICC, p, desired CI width, and, for Method 2, total population size. For Method 3, the sample sizes are different for Scenario 1 ($m_3 = 18$ for $n = 15$) as compared to Scenario 2 ($m_3 = 14$ for $n = 15$). The required sample size for Method 3 is smaller when there are many smaller clinics (Scenario 2) as compared to fewer larger clinics with the same overall population size (Scenario 1).

Scenario 3 (*moderate ICC, moderate number of variable sized clinics*): $ICC = 0.01$, $p = 0.80$, desired CI width is $\pm 5\%$ ($L = 0.05$). $N = 50$ clinics. PSU sizes were drawn from a gamma distribution (shape parameter equal to 2, scale parameter equal to 100. PSU sizes below 50 or above 1000 were discarded. The PSU sizes were then shifted by subtracting

45 from each to achieve some very small PSU sizes.) The simulated PSU sizes ranged from 5 SSUs to 955 SSUs, with an average size of 180 SSUs. $M = 8978$.

Table 5.3: Average CI width for sampling $n = 10$ clinics in Scenario 3

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 10$	$m_1 = 49$ $k_{act1} = 490$	± 0.0415	$m_2 = 46$ $k_{act2} = 460$	± 0.0426	$m_3 = 36$ $k_{act3} = 360$	± 0.0477

In this scenario, the clinic sizes vary widely. Again, Method 3 (our proposed method) results in the overall smallest sample size by a sizable amount while still achieving the desired precision.

Scenario 4 (*moderate ICC, moderate number of small clinics, low desired precision*): $ICC = 0.01$, $p = 0.80$, desired CI width is $\pm 15\%$ ($L = 0.15$). $N = 50$ clinics. each of size $M_i = 30$. $M = 1500$.

Table 5.4: Average CI width for sampling $n = 5$ clinics in Scenario 4

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 5$	$m_1 = 13$ $k_{act1} = 65$	± 0.1102	$m_2 = 12$ $k_{act2} = 60$	± 0.1170	$m_3 = 9$ $k_{act3} = 45$	± 0.1426

In this scenario, very low precision is desired $\pm 15\%$, resulting in very low sample sizes for all methods. Method 3 still achieves the desired precision with the smallest overall sample size.

Scenario 5 (*high ICC, large number of large clinics*): $ICC = 0.2$, $p = 0.80$, desired CI width is $\pm 5\%$ ($L = 0.05$). $N = 100$ clinics, each of size $M_i = 100$. $M = 10000$. *Note:* for

$n = 50$, Methods 1 and 2 return negative sample sizes, indicating that there is no feasible per-clinic sample size that would achieve the desired precision.

Table 5.5: Average CI width for sampling $n = 50$ and $n = 60$ clinics in Scenario 5

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 50$	$m_1 = n/a$ $k_{act1} = n/a$	n/a	$m_2 = n/a$ $k_{act2} = n/a$	n/a	$m_3 = 8$ $k_{act3} = 400$	± 0.0505
$n = 60$	$m_1 = 24$ $k_{act1} = 1440$	± 0.0338	$m_2 = 20$ $k_{act2} = 1200$	± 0.0349	$m_3 = 5$ $k_{act3} = 300$	± 0.0509

In this scenario, the ICC is very high (0.2). Note that Methods 1 and 2 are unable to calculate per clinic sample sizes until over $n = 50$ clinics are sampled. When $n = 60$, the difference in overall sample size between Method 2 and Method 3 is immense ($k_{act2} = 1200$ and $k_{act3} = 300$, respectively), though the differences in confidence interval width are not similarly extreme. The CI width for Method 3 is slightly above 5% for these simulations.

Scenario 6 (*zero ICC, small number of large clinics*): $ICC \approx 0$, $p = 0.80$, desired CI width is $\pm$5% ($L = 0.05$). $N = 30$ clinics, each of size $M_i = 100$. $M = 3000$. *Note:* simulated ICC is 10^{-12} because ICC must be strictly positive for beta-binomial simulations.

Table 5.6: Average CI width for sampling $n = 10$ clinics in Scenario 6

	Method 1		**Method 2**		**Method 3**	
	Design	CI Width	Design	CI Width	Design	CI Width
$n = 10$	$m_1 = 33$ $k_{act1} = 330$	± 0.0406	$m_2 = 30$ $k_{act2} = 300$	± 0.0435	$m_3 = 25$ $k_{act3} = 250$	± 0.0492

In this scenario, the ICC is effectively zero. Note that when the ICC is zero, we might expect $DEFT^2 = (1 - nm/M)$. Alternatively, Method 3 predicts $DEFT^2 = (1 - m/\overline{M})$,

which is not the design effect that would result from a simple random sample. Nonetheless, in this setting, Method 2 overestimates the sample size to achieve a confidence interval width of $\pm 5\%$, while the average confidence interval resulting from Method 3 is under $\pm 5\%$.

5.5 Discussion

We describe three methods for calculating sample sizes for probability proportional to size (PPS) two-stage clustered surveys used for national disease surveillance. In all methods we assume that we know the actual prevalence and the ICC, thus these elements must be estimated from previously available data. In Method 1, no adjustment for the effect of a finite population is made at the design stage. The finite population correction can be readily applied at the analysis stage, but the end result is overestimation of the sample size, and thus a greater cost of the survey, in order to achieve the desired precision. In Method 2, a standard adjustment for the finite population is made at the design stage. This is achieved by incorporating a finite population correction into the calculation of the effective sample size. Method 2 results in smaller sample sizes than Method 1, but in many cases the numbers are quite similar. The resulting confidence intervals tend to be narrower than planned for, but wider than the confidence intervals yielded from Method 1. Finally, we present Method 3, which adjusts for the finite population correction at the first and second stages of sampling. These corrections are incorporated into the estimate of the design effect using a formula derived in Section 2. The resulting sample sizes are smaller than those from Methods 1 and 2, especially when fewer PSUs are sampled, but the confidence intervals seem to perform well; they tend to be at or slightly smaller than the desired width.

Further, we observe that Method 1 returns the same sample size for all scenarios with the same prevalence, ICC, desired precision, and number of clinics sampled. Method 2 returns the same sample size for all scenarios with the elements listed for Method 1, plus

the same overall population size. In contrast, Method 3, our proposed method, returns different sample sizes depending on the composition of the population size. The sample size is larger for countries with few larger clinics (Scenario 1) than for countries with many smaller clinics (Scenario 2). Method 3 performs well when the clinics vary in size (Scenario 3) and when the desired precision is so low that the overall sample size is small (Scenario 4). We also evaluated the methods in scenarios with high and low ICCs. When the ICC was high (Scenario 5), Methods 1 and 2 return impossible sample sizes if the number of clinics sampled are too few, while Method 3 returns a realistic sample size that performs reasonably well. When more clinics are sampled in that particular scenario, the sample sizes for Method 3 are drastically smaller than the sample sizes for Methods 1 and 2. Finally, when the ICC is effectively zero (Scenario 6), Method 3 performs well, even though the predicted design effect does not reduce to the finite population correction of a simple random sample.

We can identify a few limitations to this method. First of all, it adds a slight layer of complexity to the sample size calculations; nonetheless, these are easily coded into user-friendly calculators using software such as Excel. Another limitation is that the finite population correction is a simplified way to analyze data sampled without replacement, and there is a wealth of literature on more technically correct methods for analyzing such data, especially for PPS sampling of PSUs (Särndal et al., 2013, chap. 4). Nonetheless, one of the major goals when prescribing methods to be used in the field is to simplify wherever possible. The simplicity of approximations vastly outweighs any risk of inefficiency or the introduction of a slight theoretical bias. Furthermore, in practice one very often uses the ubiquitous systematic sampling scheme for the sampling of PSUs, potentially with implicit stratification by geographic region; this approach does not lend itself well to standard "without replacement" methodologies (Wolter, 2007, sect. 8.6).

On the other hand, we believe that this methodology for calculating sample size for a two-stage clustered survey has many significant advantages. The greatest of these is the ability to realize cost-savings at the design stage – and, in some cases, these savings may be sig-

nificant when compared to the standard method for incorporating the finite population correction – and yet they retain their accuracy. For resource-limited small countries, this can represent immense savings. For example, for the surveillance of HIV drug resistance, each SSU sampled is a patient whose viral strain must be genotyped, which is an expensive procedure. By improving feasibility, we increase the likelihood that these surveys will actually be implemented in the first place, which is important as they can provide valuable information to national program directors. Outside of the HIV drug resistance surveillance setting, this methodology could be applied to other national surveillance activities in resource-limited settings, and the formulae can also be readily adapted for use with continuous variables or stratification.

6. Development of a viral load suppression measure adjusted for non-retention for the surveillance of acquired HIV drug resistance

Natalie Exner and Marcello Pagano

Abstract

The World Health Organization is redesigning their guidance for the surveillance of acquired HIV drug resistance in low- and middle-income countries. This paper focuses on designing such a survey. The previous survey was a longitudinal survey of HIV-infected patients during the first 12 months of treatment, but this has been replaced by a cross-sectional survey of patients on treatment for 12 ± 3 months for reasons of feasibility. There are important limitations of cross-sectional surveys. One of the key survey outcomes for this survey is the prevalence of viral load suppression (VLS, defined as viral load < 1000 copies/mL) among patients on antiretroviral therapy for 12 ± 3 months. Because the population observable from a cross-sectional survey excludes patients who have died or have been lost to follow-up, observed VLS has limited epidemiologic utility for national HIV program managers. We highlight the importance of measuring 12 month retention to assist in the interpretation of observed VLS results. In addition, we propose a novel adjusted VLS measure that incorporates data on site-specific retention by assuming that all patients who have been lost to follow-up are not virally suppressed. We believe that this adjusted VLS measure has improved utility for assessing changes in VLS over time within a country, across countries, and relative to a global standard.

6.1 Introduction

In the last ten years, low- and middle-income countries have been able to significantly expand access to antiretroviral therapy (ART) for their HIV-infected populations. Treatment efficacy is affected by HIV drug resistance (HIVDR) that is transmitted via infection or acquired via drug selective pressure. As genotyping for HIVDR mutations is prohibitively expensive in many settings, countries can implement sampling surveys to obtain nationally representative estimates of factors associated with acquired HIVDR. Acquired HIVDR (ADR) is defined as any drug resistance mutation that emerges under the selec-

tive pressure of ART (World Health Organization, 2012a). There are various ways that a patient may develop HIVDR mutations, including suboptimal adherence to treatment regimens, treatment interruption, inadequate plasma drug concentrations, or the use of suboptimal drugs or drug combinations (World Health Organization, 2012a). National program managers can use information from ADR surveys to identify gaps in service delivery and to assess the expected effectiveness of available first- and second-line regimens (World Health Organization, 2012a).

The newly revised WHO protocol for ADR surveillance in low- and middle-income countries includes a cross-sectional survey of patients on treatment for $12\pm$ 3 months. The primary outcome is the prevalence of viral load suppression (VLS, defined as viral load $<$ 1000 copies/mL) among patients retained on treatment (observable patients), and one of the secondary outcomes is the prevalence of HIVDR mutations among patients with viral suppression failure. Observed VLS is the primary outcome because it is a key indicator of program performance at 12 months. Patients who are virally suppressed do not have effective drug resistance (McMahon et al., 2013). In addition, viral suppression failure in a patient may be attributable to either the existence of HIVDR mutations or to personal or programmatic failures such as poor adherence or treatment stock-outs. These 12 month viral suppression and drug resistance outcomes provide important feedback to national HIV program managers.

The previous WHO protocol to study this issue was a longitudinal survey following a cohort of patients receiving ART during their first 12 months on treatment (World Health Organization, 2012c). The WHO abandoned the previous protocol because of the logistical complexity associated with a longitudinal survey that requires following a cohort for 12 months and the long lag between survey initiation and availability of results (World Health Organization, 2012b). In contrast to a longitudinal survey that requires a group of patients to be assessed continuously over time, a cross-sectional study only observes patients at a particular point in time; clearly, this is much less expensive than a continuously ongoing study, but a cross-sectional survey excludes patients who are no longer

receiving ART at the study site and therefore cannot be observed because they have died, been lost to follow-up or have stopped treatment. This survivor bias can significantly impact the interpretation of the primary outcome. Without accounting for within country or country-to-country variability in retention patterns, there are many important confounding factors, making it challenging to meaningfully a) assess changes in the national estimate of observed VLS over time, b) compare these estimates against a global standard, or c) compare estimates across countries.

In this paper, we discuss the epidemiologic limitations of observed VLS in the absence of complete data on retention, and we propose a new outcome that attempts to correct this bias by combining observed cross-sectional VLS with data on patient ART retention. Retention needs to be estimated using a second site-specific sample survey of patient records. This then leads to an adjusted outcome that has improved utility in that it can be more meaningfully compared across time and to a global standard. This adjusted outcome is akin to a lower bound of 12 month viral suppression because it assumes that all patients who are not retained are not virally suppressed.In our experience, the percentage of patients with documented transfer is very low in low- and middle-income countries. The treatment of documented transfers is described in Section [sub:Retention]. The primary advantage of this framework is that it measures an outcome that is very similar to that originating from a longitudinal study (i.e. prevalence of VLS among patients who initiated treatment 12 months prior) using a cross-sectional study.

We also describe how the ADR survey framework can be used to develop a nationally representative estimate of 12 month retention. Currently, 12 month retention is a recommended UNGASS/PEPFAR indicator, though its suggested implementation is via census. Using a sampling framework can drastically increase the feasibility and acceptability of this important indicator.

In Section 6.2, we provide motivation for the adjusted VLS outcome. In Section 6.3, we provide a framework and the necessary formulae for the analysis of adjusted VLS. In Sec-

tion 6.4, we provide formulae for approximating the precision of adjusted VLS to inform survey design. In Section 6.5, we describe a simulation study to evaluate the robustness of our results. In Section 6.6, we discuss the conclusions we can draw from this study.

6.2 Motivation for adjusted VLS outcome

The goal of this survey is to observe VLS patterns a year after patients have been placed on treatment. The available primary outcome of this survey is VLS among observable patients retained on ART for 12$\pm$3 months. Because this outcome is measured via a cross-sectional survey, it has important epidemiological limitations. If one only observes those patients remaining on treatment at a particular site, countries with the worst retention may appear to have the highest observed VLS because the sickest patients have been lost to follow-up or died. On the other hand, a country that makes a concerted effort to improve retention may experience a decrease in observed VLS because these newly retained patients may be failing therapy more so than those who were retained in the past. Thus, it is not meaningful to compare observed VLS over time even within the same country if retention patterns change over time[1]. Observed VLS from a cross-sectional survey provides incomplete information about program performance if a measure of retention is not incorporated into the evaluation. As a result, we measure 12 month retention in a nationally representative fashion using methodology consistent with an existing UNGASS/PEPFAR indicator (UNAIDS, 2011, sect. 4.2). Furthermore, we propose an adjusted VLS measure that incorporates information on retention; this measure makes two assumptions: (i) all patients who are lost to follow-up or die are not virally suppressed; and (ii) all patients who are documented to have transferred care to another site are assumed to have the same rate of VLS and retention as other patients.

Using a simple law of total probability, an estimate of the overall prevalence of VLS can be written as follows:

[1]This is an example of the Neyman incidence-prevalence bias

$$
\begin{aligned}
\Pr(VLS) &= \Pr(VLS|Retained)\Pr(Retained) \\
&\quad + \Pr(VLS|Not\ Retained)\Pr(Not\ Retained) \\
&= \Pr(VLS|Retained)\Pr(Retained) + 0 \\
&= \Pr(VLS|Retained)\Pr(Retained)
\end{aligned}
$$

Thus, observed VLS, i.e., $\Pr(VLS|Retained)$, and retention, i.e.,$\Pr(Retained)$, can be multiplied to yield an estimate of the overall prevalence of VLS, adjusted for non-retention. We recommend calculating adjusted VLS at the level of a site administering ART by multiplying a site-level estimate of unadjusted VLS (censoring documented transfers) and a site-level estimate of retention. These site-level adjusted VLS estimates are then combined across sites, weighting by site size, resulting in an overall estimate of VLS adjusted for non-retention. In summary, we extract those who are documented transfers and assume they behave the same as other patients. We then assume that those lost to follow-up and those who die are failures. In this way, we have imputed or measured an outcome for sampled patients who initiated therapy a year earlier. In fact, the adjusted VLS measure is very similar to one of the primary outcomes of the previous longitudinal survey, HIV drug resistance prevention (World Health Organization, 2012a, annex 1, sect. 8). HIV drug resistance prevention is defined as as the proportion of patients who initiated therapy who are virally suppressed at 12 months. Within this indicator is the inherent assumption that all patients who have been lost to follow up or died have failed. Thus, we can readily argue that adjusted VLS is measuring the same population quantity as HIV drug resistance prevention, but adjusted VLS requires only cross-sectional information.

6.3 ADR survey implementation

6.3.1 Survey design overview

The proposed survey is a two-stage clustered survey of (1) sites administering ART in the country, and (2) patients on treatment for 12±3 months during a pre-defined survey period lasting 6 months. The primary sampling units (PSUs) are the sites. The secondary sampling units (SSUs) are eligible patients.

To perform the ADR survey, n of N total sites are selected using probability proportional to size (PPS) sampling without replacement in which larger sites are thus more likely to be sampled, and then the same number of patients are included from each chosen site.[2] The sizes of the sites are estimated using available proxy information, such as the number of patients on treatment during the previous year at those sites. Because these sizes will likely differ from the actual sizes of the eligible population (i.e., patients on treatment for 12±3 months during the survey period), we refer to this type of sampling as probability proportional to proxy size (PPPS) sampling. For the $i \in 1, ..., n$ sites sampled, the PSU sampling weight is equal to the inverse of the probability that the site is selected into the sample S_I, i.e. $w_{PSU,i} = [\Pr(i \in S_I)]^{-1}$.

6.3.2 Observed VLS

Once n sites have been sampled, eligible patients are consecutively enrolled at each of these sites for the measurement of observed VLS. The target number of patients to be sampled from each site is m, as determined by the desired precision of the observed VLS outcome, but the actual number sampled may vary across sites because of differential laboratory failure or under-enrollment, for example. Among the m_i individuals sampled at site i, $\widehat{t}_i$ is the number of patients achieving VLS. Thus, the observed prevalence of VLS

[2] One such method for conducting sampling is PPS systematic sampling; the advantage of this option is an increase in ease of implementation (Wolter, 2007, sect. 8.6).

among patients retained on therapy at site i is $\widehat{p}_{VLS,i} = \widehat{t}_i / m_i$.

In order to appropriately weight the estimator, it is important that the site screen patients to determine the total number of eligible patients observed during the 6 month survey period; we refer to the observed eligible population size of site i as M_i. The SSU sampling weight, $w_{SSU,VLS,i}$, is equal to the inverse of the probability that patients from site i are selected into the sample. Since we assume that patients are randomly selected with equal probability (no time trend over the survey period), $w_{SSU,VLS,i} = M_i / m_i$. The overall sampling weight for patients with observed viral load data is $w_{VLS,i} = w_{PSU,i} w_{SSU,VLS,i}$.

A nationally representative estimate of observed VLS prevalence is a ratio with numerator equal to an estimate of the total number of patients in the country on treatment for 12±3 months who are still retained in care and are achieving viral load suppression ($\widehat{T}$); the denominator is an estimate of the total number of patients in the country on treatment for 12-24 months who are still retained in care ($\widehat{M}$). The point estimator and linearized variance estimator (with finite population corrections consistent with those in Stata's SVY command) are as follows (Lohr, 2010, eq. 6.33) (StataCorp, 2013):

$$
\begin{aligned}
\widehat{T} &= \sum_{i=1}^{n} w_{VLS,i} \widehat{t}_i \\
\widehat{M} &= \sum_{i=1}^{n} w_{VLS,i} m_i \\
\widehat{p}_{VLS} &= \frac{\widehat{T}}{\widehat{M}} \\
\widehat{var}(\widehat{p}_{VLS}) &= \left(1 - \frac{n}{N}\right) \frac{1}{\widehat{M}^2} \frac{n}{n-1} \sum_{i=1}^{n} w_{VLS,i}^2 \left(\widehat{t}_i - \widehat{p}_{VLS} m_i\right)^2 \\
&\quad + \frac{1}{\widehat{M}^2} \frac{n}{N} \sum_{i=1}^{n} \left(1 - \frac{m_i}{M_i}\right) \frac{m_i^2}{m_i - 1} w_{VLS,i}^2 \widehat{p}_{VLS,i} \left(1 - \widehat{p}_{VLS,i}\right)
\end{aligned}
$$

A 95% confidence interval can be calculated using a Wald-type interval for a proportion.

$$\widehat{p}_{VLS} \pm t_{n-1,0.975}\sqrt{\widehat{var}(\widehat{p}_{VLS})}$$

The survey analysis can be readily modified for settings when the sites are stratified prior to sampling or all sites are included in the sample StataCorp (2013).

6.3.3 Retention

In order to obtain a measure of retention for patients on treatment, we can perform a retrospective chart review at the same n sites sampled. We list patients who initiated therapy during a pre-defined sampling window, and a random sample of these patients is selected for assessment of retention. This sample can be obtained via systematic sampling (e.g., every 10th record beyond a random start point) (Lohr, 2010, sect. 2.7). 12 month retention is defined as the patient being retained on ART at exactly 12 months after treatment initiation (UNAIDS, 2011, sect. 4.2). Patients who have stopped treatment, died, or been lost to follow-up are not considered retained. Patients with documented transfer to another site are excluded from the sample. The inherent assumption is that these transferred patients have the same prevalence of retention and viral load suppression after transferring care as patients who did not transfer, and excluding them from the sample will properly implement these assumptions.

The target number of patients to be sampled from each site is s as determined by the desired precision of the estimate of 12 month retention. The number of charts reviewed may vary across sites. Among the s_i patients sampled from site i (excluding documented transfers), $\widehat{u}_i$ is the number of patients retained on treatment at 12 months. Thus, the observed prevalence of 12 month retention among patients at site i is $\widehat{p}_{RET,i} = \widehat{u}_i/s_i$.

In order to appropriately weight the estimator, it is important that the site determines the total number of eligible patients for the retrospective chart review; we refer to the total number of eligible patient records at site i as S_i. The SSU sampling weight $w_{SSU,RET,i}$

is equal to the inverse of the probability that patients from site i are selected into the sample. Since we assume that patients are randomly selected with equal probability, $w_{SSU,RET,i} = S_i/s_i$. The overall sampling weight for patients with retention data (excluding documented transfers) is $w_{RET,i} = w_{PSU,i}w_{SSU,RET,i}$.

A nationally representative estimate of 12 month retention is a ratio with numerator equal to an estimate of the total number of patients retained on treatment for 12 months who initiated during a pre-defined sampling window ($\widehat{U}$); the denominator is an estimate of the total number of patients who initiated treatment during a pre-defined sampling window ($\widehat{S}$). The point estimator and linearized variance estimator (with finite population corrections consistent with those in Stata's SVY command) are as follows (Lohr, 2010, eq. 6.33) (StataCorp, 2013):

$$\begin{aligned}
\widehat{U} &= \sum_{i=1}^{n} w_{RET,i}\widehat{u}_i \\
\widehat{S} &= \sum_{i=1}^{n} w_{RET,i}s_i \\
\widehat{p}_{RET} &= \frac{\widehat{U}}{\widehat{S}} \\
\widehat{var}(\widehat{p}_{RET}) &= \left(1-\frac{n}{N}\right)\frac{1}{\widehat{S}^2}\frac{n}{n-1}\sum_{i=1}^{n} w_{RET,i}^2\left(\widehat{u}_i - \widehat{p}_{RET}s_i\right)^2 \\
&\quad + \frac{1}{\widehat{S}^2}\frac{n}{N}\sum_{i=1}^{n}\left(1-\frac{s_i}{S_i}\right)\frac{s_i^2}{s_i-1}w_{RET,i}^2\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)
\end{aligned}$$

A 95% confidence interval can be calculated using a Wald-type interval for a proportion.

$$\widehat{p}_{RET} \pm t_{n-1,0.975}\sqrt{\widehat{var}(\widehat{p}_{RET})}$$

The survey analysis can be readily modified for settings when the sites are stratified prior to sampling or all sites are included in the sample (StataCorp, 2013).

6.3.4 Adjusted VLS

At each of the n sites sampled for the survey, we can estimate the site-specific adjusted VLS prevalence as the product of site-specific observed VLS and site-specific 12 month retention. We then weight the site results by the number of patients who initiated therapy at each site (S_i for site i). The adjusted VLS measure can be estimated as a ratio with numerator being an estimate of the total number of patients in the country retained on treatment for 12 months with viral load suppression ($\widehat{V}$); the denominator is an estimate of the total number of patients who initiated treatment during a pre-defined sampling window ($\widehat{S}$). (*Note*: this is equivalent to $\widehat{S}$ calculated for the retention denominator.) The point estimator and linearized variance estimator (with finite population corrections consistent with those in Stata's SVY command) are as follows (see Appendix A.2.1):

$$
\begin{aligned}
\widehat{V} &= \sum_{i=1}^{n} w_{PSU,i} S_i \hat{p}_{VLS,i} \hat{p}_{RET,i} \\
\widehat{S} &= \sum_{i=1}^{n} w_{PSU,i} S_i \\
\widehat{p}_{ADJ} &= \frac{\widehat{V}}{\widehat{S}} \\
\widehat{var}(\widehat{p}_{ADJ}) &= \frac{1}{\widehat{S}^2}\left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\right)^2\left(\widehat{p}_{VLS,i}\widehat{p}_{RET,i}-\widehat{p}_{ADJ}\right)^2 \\
&\quad +\frac{1}{\widehat{S}^2}\frac{n}{N}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\right)^2\left\{\widehat{p}_{VLS,i}^2\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)}{s_i}\right. \\
&\quad +\widehat{p}_{RET,i}^2\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i} \\
&\quad \left.-\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i}\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)}{s_i}\right\}
\end{aligned}
$$

A 95% confidence interval can be calculated using a Wald-type interval for a proportion.

$$\widehat{p}_{ADJ} \pm t_{n-1,0.975}\sqrt{\widehat{var}(\widehat{p}_{ADJ})}$$

6.4 ADR survey design

Prior to implementing the survey, countries must determine a suitable design that will be a compromise between desired precision and logistical/financial feasibility. Guidance on how to calculate sample size requirements to achieve a certain precision, for the observed prevalence of VLS, $\widehat{p}_{VLS}$, is described elsewhere (see Chapter 5). The same procedure can be used to calculate sample size requirements to achieve a certain precision for the retention measure, $\widehat{p}_{RET}$. Given the sample sizes identified for the viral load suppression and retention portions of the survey, it is useful for countries to predict the precision of the adjusted VLS measure, $\widehat{p}_{ADJ}$.

With some assumptions, we derive estimates of the variance of the adjusted VLS measure that can be used to predict the expected confidence interval width resulting from a survey with a particular combination of sample sizes for VLS and retention. These approximations assume that site size, site-specific prevalence of observed VLS, and site-specific retention are independent. Sensitivity to these assumptions is evaluated in our simulations in Section 6.5. For the derivations and simulations, we assume probability proportional to size (PPS) sampling, in which the actual site sizes are known, although we acknowledge that generally this information will not be available prior to site sampling. In this case, the predicted variances below should be multiplied by an additional design effect due to disproportionate weighting, often expressed as $1 + cv^2(weights)$, where $cv(\cdot)$ indicates the coefficient of variation (Park and Lee, 2004, eq. 2.2).

For an infinite population, the variance of $\widehat{p}_{ADJ}$ can be approximated using Equation 6.1, requiring the parameters for the survey design (n sites sampled, m VLS patients sampled per clinic, and s retention records sampled per clinic), an estimate of the prevalence of observed VLS (p_{VLS}), an estimate of the intracluster correlation coefficient for observed VLS (ICC_{VLS}), an estimate of the prevalence of retention (p_{RET}), and an estimate of the intracluster correlation coefficient for retention (ICC_{RET}). Note that all of these elements are already required for the design and implementation of the surveys for observed VLS

($\widehat{p}_{VLS}$) and retention ($\widehat{p}_{RET}$). When a subset of sites are sampled in the absence of stratification, the predicted variance is as follows (see Appendix A.2.2):

$$\begin{aligned} Var(\widehat{p}_{ADJ}) \approx \frac{1}{n}\Bigg\{ & \left[ICC_{VLS} + \frac{1}{m}\left(1 - ICC_{VLS}\right)\right] p_{RET}^2 p_{VLS}(1-p_{VLS}) \\ & + \left[ICC_{RET} + \frac{1}{s}\left(1 - ICC_{RET}\right)\right] p_{VLS}^2 p_{RET}(1-p_{RET}) \\ & + \left[ICC_{VLS} + \frac{1}{m}\left(1 - ICC_{VLS}\right)\right]\left[ICC_{RET} + \frac{1}{s}\left(1 - ICC_{RET}\right)\right] \\ & \times p_{VLS}(1-p_{VLS}) p_{RET}(1-p_{RET})\Bigg\} \end{aligned} \tag{6.1}$$

For a finite population with N total sites, $M = \sum_{i=1}^{N} M_i$ total patients eligible for viral load suppression testing, and $S = \sum_{i=1}^{N}$ total records eligible for retention estimation, the variance of $\widehat{p}_{ADJ}$ can be approximated by Equation (see Appendix A.2.3):

$$\begin{aligned} Var(\widehat{p}_{ADJ}) \approx \frac{1}{n}\Bigg\{ & \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{N}{M}\right)\left(1 - ICC_{VLS}\right)\right] p_{RET}^2 p_{VLS}(1-p_{VLS}) \\ & + \left[ICC_{RET} + \left(\frac{1}{s} - \frac{N}{S}\right)\left(1 - ICC_{RET}\right)\right] p_{VLS}^2 p_{RET}(1-p_{RET}) \\ & + \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{N}{M}\right)\left(1 - ICC_{VLS}\right)\right] \\ & \times \left[ICC_{RET} + \left(\frac{1}{s} - \frac{N}{S}\right)\left(1 - ICC_{RET}\right)\right] \\ & \times p_{VLS}(1-p_{VLS}) p_{RET}(1-p_{RET})\Bigg\} \end{aligned} \tag{6.2}$$

The corresponding predicted confidence interval half-width using a Wald-type interval for a proportion is:

$$t_{n-1,0.975}\sqrt{Var(\widehat{p}_{ADJ})}$$

6.5 Simulations

6.5.1 Simulation set-up

We evaluate these methods by simulating data within a hypothetical large country and a hypothetical small country. In both cases, we assume that the national prevalence of VLS among retained patients is $p_{VLS} = 85\%$, and the national prevalence of 12-month retention is $p_{RET} = 85\%$. Site-specific prevalence of VLS, $p_{VLS,i}$, is drawn from a $Beta(\alpha_{VLS},\ \beta_{VLS})$ for $i = 1, ..., N$ sites in the country. The parameters of the beta distribution are the following, where $ICC_{VLS} = 0.0043$ is the intracluster correlation of observed VLS:

$$\begin{aligned} \alpha_{VLS} &= \left[\frac{1 - ICC_{VLS}}{ICC_{VLS}}\right] p_{VLS} \\ \beta_{VLS} &= \left[\frac{1 - ICC_{VLS}}{ICC_{VLS}}\right] (1 - p_{VLS}) \end{aligned}$$

Similarly, site-specific retention, $p_{RET,i}$, is drawn from a $Beta(\alpha_{RET},\ \beta_{RET})$ for $i = 1, ..., N$ sites in the country. The parameters of the beta distribution are the following, where $ICC_{RET} = 0.0476$ is the intracluster correlation of 12-month retention:

$$\begin{aligned} \alpha_{RET} &= \left[\frac{1 - ICC_{RET}}{ICC_{RET}}\right] p_{RET} \\ \beta_{RET} &= \left[\frac{1 - ICC_{RET}}{ICC_{RET}}\right] (1 - p_{RET}) \end{aligned}$$

The assumed values described above are identical to those used in the proposed ADR surveillance guidance. Their justification is described elsewhere (see Chapter 4).

To simulate the patient outcomes of the S_i eligible records that can be reviewed for retention in a site, we assign $M_i = S_i p_{RET,i}$ as retained, and we assign the remaining $S_i(1 - p_{RET,i})$ as not retained. Among the M_i retained patients, we assign $M_i p_{VLS,i}$ as virally suppressed, and we assign the remaining $M_i(1 - p_{VLS,i})$ as not virally suppressed.

To simulate two stage cluster sampling, we randomly sample n of N sites using probability proportional to size (PPS) sampling without replacement; sampling is proportional to the number of eligible records at each site, S_i. For the observed VLS outcome, we randomly assess the VLS outcomes of m retained patients from each selected site using simple random sampling without replacement. If m is larger than the number of retained patients in the selected site, sampling stops after all available patients are included. Sampling for the retention outcome proceeds similarly, with s patient records being sampled from each selected site. For the purposes of this simulation, m and s are selected to achieve confidence intervals of $\pm$5% around $\widehat{p}_{VLS}$ and $\widehat{p}_{RET}$ (see Chapter 4).

For each simulated cluster sampled, we estimate the adjusted VLS prevalence, standard error, and 95% confidence interval as using the formulas provided in Section 6.3.4. We repeat the simulations 10,000 times. We calculate the true value of the adjusted VLS outcome, which is $p_{ADJ} = \sum_{i=1}^{N} S_i p_{VLS,i} p_{RET,i} \Big/ S$. We report the average point estimate from the 10,000 simulations. We calculate the simulation standard error, which is the standard deviation of the 10,000 simulated prevalence estimates. We report the average estimated standard error and average confidence interval (CI) width. The average confidence interval width should be compared to the predicted CI width as calculated by Equation 6.2.

Unlike the proposed ADR surveillance guidance, we do not incorporate adjustments for laboratory failure, the proportion of patients on first-line regimens, documented transfer, and disproportionate weighting as these are not features of the simulation. Details of how to accommodate these design features is described elsewhere (see Chapter 4). Briefly, the design effect is multiplied by an inflation factor to account for disproportionate weighting, and the sample size is inflated by dividing by the expected proportion of eligible patients (eg. divide by 0.90 if 10% laboratory failure is expected).

6.5.2 Large country

We simulate a large country, with $N = 2000$ sites and an average of approximately 200 eligible records per site. This hypothetical country is intended to be similar to large countries in Sub-Saharan Africa with generalized epidemics. The site sizes are sampled from a truncated gamma distribution scaled to have mean 200; simulated sizes ranged from approximately 40 to 1000 patients per site. The proposed design requires sampling $n = 20$ sites, $m = 12$ observed patients for VLS assessment, and $s = 21$ patient records for retention assessment. The predicted CI width by Equation 6.2 is $\pm$5.87%.

To challenge the assumptions of the derivations, data are simulated in a variety of ways. (1) Site size (S_i), site-specific observed VLS ($p_{VLS,i}$), and site-specific retention ($p_{RET,i}$) are mutually independent. (2) $p_{RET,i}$ is independent of $p_{VLS,i}$ and S_i, but S_i and $p_{VLS,i}$ are sorted so they have perfect positive rank correlation; this corresponds to larger sites having better VLS among retained patients. (3) Same as setting (2), except S_i and $p_{VLS,i}$ have perfect negative rank correlation; this corresponds to smaller sites having better VLS among retained patients. (4) $p_{VLS,i}$ is independent of $p_{RET,i}$ and S_i, but S_i and $p_{RET,i}$ are sorted so they have perfect positive rank correlation; this corresponds to larger sites having better patient retention. (5) Same as setting (4), except S_i and $p_{RET,i}$ have perfect negative rank correlation; this corresponds to smaller sites having better retention. (6) S_i is independent of $p_{VLS,i}$ and $p_{RET,i}$, but $p_{VLS,i}$ and $p_{RET,i}$ are sorted so they have perfect positive rank correlation; this corresponds to sites with higher VLS having higher retention. (7) Same as setting (6), except $p_{VLS,i}$ and $p_{RET,i}$ have perfect negative rank correlation; this corresponds to sites with higher VLS having lower retention.

The estimator has little to no bias in the settings we investigate. The estimated standard error using the variance estimator in Section 6.3.4 is close to the simulated standard error in all settings, but it tends to be a slight underestimate. The average CI width is roughly similar to the predicted CI width in all settings. The average CI width is especially large in two settings; these settings are when site size is negatively correlated with retention

Table 6.1: Large country simulation results

	Truth p_{ADJ}	Estimated $\widehat{p}_{ADJ}$	Simulation SE $Var(\widehat{p}_{ADJ})$	Estimated SE $\widehat{var}(\widehat{p}_{ADJ})$	Average CI Width Predicted ±5.87%
(1) $(S_i \perp p_{VLS,i} \perp p_{RET,i})$	0.716	0.716	0.0282	0.0281	±5.89%
(2) $(+S_i, +p_{VLS,i})$	0.727	0.727	0.0284	0.0280	±5.86%
(3) $(+S_i, -p_{VLS,i})$	0.704	0.704	0.0291	0.0286	±5.98%
(4) $(+S_i, +p_{RET,i})$	0.754	0.754	0.0270	0.0266	±5.56%
(5) $(+S_i, -p_{RET,i})$	0.673	0.673	0.0300	0.0295	±6.18%
(6) $(+p_{VLS,i}, +p_{RET,i})$	0.716	0.716	0.0308	0.0301	±6.30%
(7) $(+p_{VLS,i}, -p_{RET,i})$	0.716	0.716	0.0269	0.0262	±5.47%

(Setting 5) and when observed VLS and retention are positively correlated (Setting 6). The average CI width is especially low in two settings; these settings are when site size is positively correlated with retention (Setting 4) and when observed VLS and retention are negatively correlated (Setting 7). When retention is independent of both site size and observed VLS (Settings 1, 2, and 3), the predicted CI width performs well.

6.5.3 Small country

We simulate a small country, with $N = 50$ sites and an average of approximately 20 eligible records per site. This hypothetical country is intended to be similar to small countries in Latin America with concentrated epidemics. The site sizes are sampled from a truncated gamma distribution scaled to have mean 20; simulated sizes range from approximately 4 to 90 patients per site. The proposed design requires sampling $n = 20$ sites, $m = 8$ observed patients for VLS assessment, and $s = 9$ patient records for retention assessment. (Note that small sites will under-enroll patients.) The predicted CI width by Equation 6.2 is ±6.01%.

We observe a slight bias in the mean point estimate, with the maximal bias being 2.0% (Setting 5). For all settings, the estimated standard error using the variance estimator in Section 6.3.4 is close to the simulated standard error, but it tends to underestimate the

Table 6.2: Small country simulation result

	Truth p_{ADJ}	Estimated $\widehat{p}_{ADJ}$	Simulation SE $Var(\widehat{p}_{ADJ})$	Estimated SE $\widehat{var}(\widehat{p}_{ADJ})$	Average CI Width Predicted ±6.01%
(1) $(S_i \perp p_{VLS,i} \perp p_{RET,i})$	0.700	0.701	0.0281	0.0279	±5.84%
(2) $(+S_i, +p_{VLS,i})$	0.720	0.720	0.0278	0.0276	±5.79%
(3) $(+S_i, -p_{VLS,i})$	0.697	0.703	0.0283	0.0281	±5.89%
(4) $(+S_i, +p_{RET,i})$	0.750	0.742	0.0257	0.0254	±5.33%
(5) $(+S_i, -p_{RET,i})$	0.659	0.679	0.0281	0.0284	±5.94%
(6) $(+p_{VLS,i}, +p_{RET,i})$	0.712	0.710	0.0286	0.0282	±5.91%
(7) $(+p_{VLS,i}, -p_{RET,i})$	0.714	0.713	0.0262	0.0258	±5.41%

standard error slightly more than in the large country setting. The average CI width is below the predicted width of ±6.01% in all settings. The mean confidence intervals have similar trends in relative width as described for the hypothetical large country (i.e., widest for Settings (5) and (6), narrowest for Settings (4) and (7), and moderate for Settings (1), (2), and (3).

6.6 Discussion

The measurement of VLS among HIV-infected patients retained on ART for 12±3 months using a cross-sectional study has severe epidemiological limitations. It excludes patients who have died or been lost to follow-up since ART initiation. Thus, because of the confounding, it is not meaningful to compare measures of observed VLS over time or across regions if mortality or retention rates have changed. We describe the importance of collecting data on retention and incorporating it into the calculation of what is a key indicator of national HIV program performance. We describe an approach for developing a more epidemiologically meaningful measure of VLS in HIV-infected patients. In this approach, representative data on retention is combined with representative data on observed VLS. For patients who die or are lost to follow-up, we assume that they are not virologically suppressed. We refer to this measure as adjusted VLS, although it is akin to a lower-bound

of VLS.

We develop formulae that can be used to approximate the variance of adjusted VLS under a set of simplifying assumptions (Section 6.4). These formulae assume PPS sampling, and they should be supplemented by an additional design effect factor to account for the expected disproportionate weighting due to imperfect prior information on site size (see Chapter 4, suggests multiplying by 1.50). These formulae can assist program managers when they are determining the necessary sample sizes for the VLS and retention portions of the ADR survey.

In Section 6.5, we describe our simulation study, assessing this outcome in a hypothetical large and a hypothetical small country under a variety of different assumptions. In small countries, the point estimate can have slight bias. This is not surprising as the point estimate is calculated as a ratio estimate, which is known to be slightly biased but is used because of its lower mean square error (Lohr, 2010, sect. 4.1.2). The standard error estimator tends to slightly underestimate the variability, but it seems to work well in all of the different simulated settings, regardless of country size and of the correlation between site size, prevalence of observed VLS, and retention. The predicted confidence interval width is similar to the observed confidence interval width even when some of the assumptions are severely violated.

We can identify a few limitations to this method. It makes the strong assumption that all patients who are not retained on therapy are not virologically suppressed. In the absence of data on missing patients, this is likely the most reasonable assumption available. Data from a study in urban Malawi reported, among lost patients, 30% had died (Tweya et al., 2013). Among those who were still alive, 44% had stopped taking ART entirely. Thus, for these segments of the population, assuming that these patients are not virologically suppressed (with death defined as a virological failure) seems reasonable. On the other hand, for the remaining 56% of living patients who reported still taking ART by sourcing drugs from other sites, using alternative ART sources, or making brief ART interruptions,

the assumption that none of these patients are suppressed may be overly pessimistic. This highlights the limitation that the proposed method is unable to detect undocumented, or 'silent,' transfers out of the site. This limitation stems from the larger challenge of properly estimating retention in resource-limited settings, and it is also a limitation of the previously used longitudinal survey. Thus, while it affects the interpretation of this measure, it is not a failure of the method itself. An additional limitation is the increased complexity of the analysis formulae and the approximation formulae, although these can be readily coded into any statistical software package or spreadsheet-based program.

Overall, we believe that this adjusted VLS measure has increased epidemiological utility over the observed, on-site VLS measure, and, in fact, it is very similar to the HIV drug resistance prevention measure previously described for the longitudinal acquired drug resistance protocol. To achieve a VLS measure with improved utility, we highlight the importance of collecting representative data on retention for the interpretation of acquired drug resistance outcomes in patients on therapy for a fixed amount of time. The benefit of the adjusted VLS measure is that it fits into the existing design-based survey analysis framework, requiring only small modifications to the variance estimator. This measure can be calculated from survey data collected on observed VLS and retention at no additional cost. Because it can be compared within countries over time, across countries, and to a global standard, it is an important and useful addition to the acquired HIV drug resistance surveillance guidance.

7. Evaluating confidence interval methods for binomial proportions in clustered surveys

Natalie Exner and Marcello Pagano

Abstract

In survey settings, a variety of methods for constructing confidence intervals for proportions are available; these methods include the standard Wald method, a class of modified methods that replace the sample size with the survey effective sample size (Wilson, Clopper-Pearson, Jeffreys, and Agresti-Coull), and transformed methods (Logit and Arcsine). We describe these seven methods, two of which have not been previously evaluated in the literature (the modified Jeffreys and Agresti-Coull intervals). For each method, we describe a formulation that does and does not adjust for the design degrees of freedom. We suggest a definition of adjusted effective sample size that induces equivalency between different confidence interval expressions. We also expand on an existing framework for truncation that can be used when data appears to be more efficient than a simple random sample or when data has zero standard error and/or a point estimate of 0 or 1. We compare these methods using a simulation study modeled after the 30x7 design for immunization surveys. Our results confirmed the importance of adjusting for the design degrees of freedom. As expected, the Wald interval performed very poorly, frequently failing to achieve the nominal coverage level. For similar reasons, we do not recommend the use of the Arcsine interval. When the intracluster correlation coefficient is high and the prevalence $p < 10\%$ or $> 90\%$, the Agresti-Coull and Clopper-Pearson intervals perform best. In other settings, the Clopper-Pearson interval is unnecessarily wide. In general, the Logit, Wilson, Jeffreys and Agresti-Coull intervals perform well, though the Logit interval can be too wide. The Wilson interval performed best when a non-unimodal distribution was assumed for the simulations.

7.1 Introduction

A very important and useful inferential tool is the confidence interval. If the outcome of interest is discrete, a complication arises when one attempts to calculate an exact pre-

determined level of confidence (see Brown, Cai, and DasGupta, 2001, for an overview). When dealing with surveys, the data analysis must adjust for any clustering, stratification, or weighting used in the design. The standard and popular 95% confidence interval for a proportion in a survey setting is a Wald-based interval. In the independent and identically distributed (IID) setting, it has been extensively demonstrated that the Wald interval performs poorly for proportions, especially when the proportion is close to 0 or 1 and/or the sample size is small (Agresti and Coull, 1998; Brown et al., 2001). Coverage can be below the nominal 95% level. Also, the Wald interval can have limits below 0 or above 1, which is inappropriate for a proportion.

In the IID setting, other confidence interval methods for proportions exhibit more desirable qualities. Intervals constructed using the Wilson (quadratic), Jeffreys (beta binomial) or Clopper-Pearson (binomial) methods cannot have limits outside of the 0 to 1 range. Also, these intervals, along with the Agresti-Coull interval (modified Wald), tend to have coverage closer to the nominal 95% level, though the Clopper-Pearson interval can be unnecessarily conservative. The general conclusion is that the Wilson and Jeffreys intervals provide the best balance of confidence interval width and coverage, with the Agresti-Coull interval also performing well when the sample size is sufficiently large (Brown et al., 2001).

It is reasonable to infer that the Wald interval would also perform poorly in the complex survey setting when the expected proportion is close to 0 or 1 and/or the sample size is small, and this has been demonstrated in a few simulation studies (Korn and Graubard, 1998; Sukasih and Jang, 2005; Feng and Sitter, 2008). A variety of alternative methods for confidence interval construction for proportions have been described, including modifications of standard methods replacing the sample size with the survey effective sample size or transforming the proportion to a different scale. Literature on the topic is largely limited to conference proceedings (Kott et al., 2001; Curtin et al., 2006; Rust and Hsu, 2007; Feng and Sitter, 2008) and a publicly available masters thesis (Feng, 2006), with very few reports appearing in peer-reviewed journals (Kott and Carr, 1997; Korn and Graubard,

1998; Gray et al., 2004, being notable exceptions). We intend to remedy this situation in this paper where we assess a number of methods, including some that have not been previously applied either in simulation studies or in practice.

In addition, we discuss the existing framework for adjusting these intervals for the survey design degrees of freedom. The design degrees of freedom, denoted df_{design}, is traditionally equal to the number of primary sampling units minus the number of strata (Korn and Graubard, 1999, p. 62). While there is no formal theoretical justification for the design degrees of freedom, empirical evidence from the Wald interval suggests that accounting for this design feature can improve the performance of the confidence interval (Korn and Graubard, 1998). For each interval method in this paper, we describe alternative formulations with and without a degrees of freedom adjustment. We describe a novel method for incorporating the design degrees of freedom into the effective sample size that induces equivalency between different expressions. We also provide an in-depth discussion of truncation, a procedure by which data expected to be no more efficient than a simple random sample is truncated so that the design effect is equal to 1. We describe our recommendations for handling truncation and the related concept of 'degenerate data to improve logical consistency in the framework.

The goal of this paper is to summarize the methods available for confidence interval construction for proportions in complex surveys, study approaches for incorporating the design degrees of freedom into interval construction, describe a logically consistent framework for data truncation, evaluate the methods using a simulation study based on the popular 30x7 survey design, and provide practical guidance based on the performance of the intervals. In Section 7.2, we describe the confidence interval methods considered. In Section 7.3, we discuss our framework for truncation and handling 'degenerate' data. In Section 7.4, we describe our simulation study and the results. In Section 7.5, we discuss our conclusions.

7.2 Confidence interval methods

7.2.1 Method categories

We divide the confidence interval methods we evaluate into three categories. The first category includes only the Wald method. The second category is a class of modified methods, in which the sample size is replaced by the survey effective sample size; the survey effective size is related to the design effect, which quantifies the departure from the ideal of a simple random sample; this category includes the modified Wilson, Clopper-Pearson, Jeffreys, and Agresti-Coull intervals. The third category considers methods in which the interval is constructed on a different scale, using a Wald-type method, and then the endpoints are back-transformed to yield the final interval; this category includes the Logit and Arcsine transformations. All intervals described can be calculated from basic elements produced by typical statistical output, including the appropriately weighted point estimate, $\widehat{p}$, and the estimated standard error, $\widehat{SE}(\widehat{p})$.

7.2.2 Design degrees of freedom

For each method, we describe two alternative formulations: one that does and one that does not adjust for the design degrees of freedom. Adjustments for the design degrees of freedom can be made by replacing the standard normal quantile with a t-distribution quantile with df_{design} degrees of freedom. Alternatively, the effective sample size can be replaced by the degrees-of-freedom adjusted effective sample size, described below.

A key element for modifying IID confidence interval methods for the survey setting is calculating the survey effective sample size. The effective sample size is the sample size of a simple random sample that would yield the same precision as the survey under consideration. The effective sample size reflects the either gain or loss of precision attributable

to the survey design. The effective sample size is:

$$n_{eff} = \frac{\widehat{p}(1-\widehat{p})}{[\widehat{SE}(\widehat{p})]^2} \tag{7.1}$$

Korn and Graubard (1998) suggest the use of the degrees-of-freedom adjusted effective sample size (henceforth referred to as simply the adjusted effective sample size). The adjusted effective sample size is equal to the effective sample size multiplied by a deflation factor that reflects the difference between the actual sample size, n_{act}, and the design degrees of freedom. When constructing a two-sided confidence interval of level $1-\alpha$, the adjusted effective sample size suggested by Korn and Graubard (1998) is:

$$n^*_{eff,KG} = n_{eff}\left\{\frac{t_{n_{act}}(1-\alpha/2)}{t_{df_{design}}(1-\alpha/2)}\right\}^2$$

where $t_df(p)$ indicates the pth quantile of the t distribution with df degrees of freedom, and n_{eff} is defined as in Equation 7.1. Heuristically we argue that for finite samples the distribution of the estimator, although asymptotically normal, may be better approximated by a t distribution. We recommend using a slightly different formula for the adjusted degrees of freedom:

$$n^*_{eff} = n_{eff}\left\{\frac{z(1-\alpha/2)}{t_{df_{design}}(1-\alpha/2)}\right\}^2$$

where $z(p)$ indicates the pth quantile of the standard normal distribution. The motivation for using this formula will be described in greater detail later.

7.2.3 Standard method

The first method described is the standard Wald-type interval, also called the normal approximation or linear method. This method, which is based on the normal approximation to the binomial distribution, produces a confidence interval that is symmetric around the

point estimate, and it can produce intervals with endpoints below 0 or above 1. For a 95% confidence interval, with $\widehat{p}$ equal to the appropriately weighted point estimate, and $\widehat{SE}(\widehat{p})$ equal to the estimated standard error, the confidence interval is calculated as:

$$\hat{p} \pm z(1-\alpha/2)\widehat{SE}(\widehat{p}) \tag{7.2}$$

We can show that the above interval is equivalent to:

$$\widehat{p} \pm z(1-\alpha/2)\sqrt{\widehat{p}(1-\widehat{p})/n_{eff}} \tag{7.3}$$

This representation in expression 7.3 resembles the IID Wald method with the sample size replaced by the effective sample size. We refer to either of these confidence interval methods as the Wald method.

To adjust for the design degrees of freedom, one can replace $z(1-\alpha/2)$ in expression 7.2 with $t_{df_{design}}(1-\alpha/2)$. Alternatively, one can replace n_{eff} in expression 7.3 with n_eff^*. These intervals can be shown to be equivalent. We refer to either of these confidence interval methods as the Wald, adj. method. Note that these intervals are not equivalent if the Korn and Graubard adjusted effective sample size ($n^*_{eff,KG}$) is used instead. Thus, our proposed version of the adjusted effective sample size induces consistency between the expressions.

7.2.4 Wilson method

The first modified method we describe is the modified (or ad hoc) Wilson, suggested by Kott and Carr (1997). For the IID setting, the Wilson interval, also known as the score or quadratic interval, is constructed by solving the following quadratic function (Wilson, 1927):

$$|\widehat{p}-p|^2 \leq z(1-\alpha/2)^2\,[p(1-p)/n]$$

This is equivalent to utilizing the asymptotic distribution of $\widehat{p}$ to set the interval. To modify the interval for the survey setting, we can replace the sample size n by the effective sample size n_{eff} to produce the following formula for the upper and lower bounds of the interval:

$$\frac{\widehat{p} + \frac{z(1-\alpha/2)^2}{2n_{eff}} \pm z(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_{eff}} + \frac{z(1-\alpha/2)^2}{\left(2n_{eff}\right)^2}}}{1 + \frac{z(1-\alpha/2)^2}{n_{eff}}} \tag{7.4}$$

We refer to this method as the Wilson method.

To adjust for the design degrees of freedom, Kott and Carr (1997) suggest replacing $z(1-\alpha/2)$ in expression 7.4 with $t_{df_{design}}(1-\alpha/2)$. Alternatively, one can replace n_{eff} in expression 7.4 with n^*_{eff}. These intervals can be shown to be equivalent. We refer to either of these confidence interval methods as the Wilson, adj. method. As before, these adjusted intervals are not equivalent if the Korn and Graubard formulation of the adjusted effective sample size ($n^*_{eff,KG}$) is used.

7.2.5 Clopper-Pearson method

The second modified method adapts the traditional Clopper-Pearson interval, also called the binomial or exact interval. The Clopper-Pearson interval uses the binomial distribution, and in the IID setting, the coverage of the Clopper-Pearson interval is always at or above the nominal confidence level (Clopper and Pearson, 1934). The limits of the Clopper-Pearson interval can be calculated using the quantiles of an F or beta distribution. Using the notation of Korn and Graubard (1998), the lower and upper limits in the IID setting are defined as:

$$p_L = \frac{\nu_1 F_{\nu_1,\nu_2}(\alpha/2)}{\nu_2 + \nu_1 F_{\nu_1,\nu_2}(\alpha/2)}$$
$$p_U = \frac{\nu_3 F_{\nu_3,\nu_4}(1-\alpha/2)}{\nu_4 + \nu_3 F_{\nu_3,\nu_4}(1-\alpha/2)}$$

where $\nu_1 = 2x$, $\nu_2 = 2(n-x+1)$, $\nu_3 = 2(x+1)$, and $\nu_4 = 2(n-x)$, n is the sample size, x is the observed number of successes, and $F_{num,den}(p)$ is the pth quantile of an F distribution with num and den degrees of freedom. To construct the modified Clopper-Pearson interval, the sample size n is replaced by the adjusted effective sample size n^*_{eff}, and the observed number of successes is replaced by $\widehat{p}n^*_{eff}$. The adjusted effective sample size is used because the Clopper-Pearson interval does not otherwise account for the design degrees of freedom. We refer to this method as the CP, adj. method.

To maintain consistency with the rest of the paper, we also evaluate an unadjusted method, though no such method appears in the literature. For the unadjusted method, the sample size is replaced by the effective sample size n_{eff}, and the observed number of successes is replaced by $\widehat{p}n^*_{eff}$. We refer to this method as the CP method.

7.2.6 Jeffreys method

The next modified method is the Jeffreys interval, which is constructed from a non-informative $Beta(0.5, 0.5)$ prior for binomially distributed data (Brown et al., 2001). In the IID setting, the Jeffreys interval can be regarded as a mid-p version of the Clopper-Pearson interval, and its bounds are always contained within the bounds of the Clopper-Pearson interval, making it less conservative than the Clopper-Pearson (Brown et al., 2001). In the IID setting, the lower and upper limits are defined as:

$$p_L = Beta_{\alpha_1,\beta_1}(\alpha/2)$$
$$p_U = Beta_{\alpha_1,\beta_1}(1-\alpha/2)$$

where $\alpha_1 = x + 0.5$, $\beta_1 = n - x + 0.5$, and $Beta_{shape_1,shape_2}(p)$ is the pth quantile of a Beta distribution with $shape_1$ and $shape_2$ degrees of freedom. To modify the Jeffreys interval, the sample size n is replaced by the adjusted effective sample size n_{eff}, and the observed number of successes x is replaced by $\widehat{p}n^*_{eff}$. This modification was first suggested in the appendix of a paper (Gray et al., 2004). We refer to this as the Jeffreys method.

Alternatively, the sample size can be replaced by the adjusted effective sample size because the design degrees of freedom are not incorporated otherwise (Curtin et al., 2006). In other words, the sample size n is replaced by the adjusted effective sample size n^*_{eff}, and the observed number of successes is replaced by $\widehat{p}n^*_{eff}$. We refer to this method as the Jeffreys, adj. method. To our knowledge, neither the modified Jeffreys nor the adjusted modified Jeffreys intervals for surveys have been applied in simulations or in practice.

7.2.7 Agresti-Coull method

The Agresti-Coull method was developed to have the simplicity of the Wald interval but with performance more like the Wilson interval (Agresti and Coull, 1998). In their paper, the authors demonstrate that the midpoint for a Wilson interval is a weighted average of the observed prevalence and $1/2$. To modify the Wald interval to more closely resemble the Wilson interval, a constant is added to the number of successes and two times that constant is added to the number of trials. To construct the interval in the IID setting, the bounds are as follows:

$$\tilde{p} \pm z(1-\alpha/2)\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

where $\tilde{x} = x + c$, $\tilde{n} = n + 2c$, and $\tilde{p} = \tilde{x}/\tilde{n}$. For a 95% confidence interval, the authors suggest letting $c = 1.96^2/2 = 1.92$, but they propose that setting $c = 2$ may be easier to understand by non-statisticians because it is akin to adding two successes and two failures to the data. In this paper, we use the former, more theoretically-motivated definition

in which $c = 1.92$. Note that the Agresti-Coull interval can have bounds that are below 0 or above 1.

Similar to the other modifications, the Agresti-Coull interval can be adapted to the survey setting by letting $\tilde{x} = \widehat{p}n_{eff} + c$, $\tilde{n} = n_{eff} + 2c$. We refer to this as the AC method. Though this modification has been suggested previously (Curtin et al., 2006), it has not been applied in practice or evaluated in a simulation study, to the best of our knowledge.

The Agresti-Coull method can be further modified to adjust for the design degrees of freedom by letting $\tilde{x} = \widehat{p}n^*_{eff} + c$ and $\tilde{n} = n^*_{eff} + 2c$. We refer to this as the AC, adj. method. Note that this is not equivalent to constructing the interval using a t-quantile, $t_{df_{design}}(1 - \alpha/2)$, in place of the Z-quantile, $z(1 - \alpha/2)$. The two methods will only be equivalent if we use the Z-quantile and let $\tilde{x} = \widehat{p}n^*_{eff} + c^*$ and $\tilde{n} = n^*_{eff} + 2c^*$, where $c^* = c\left\{\frac{z(1-\alpha/2)}{t_{df_{design}}(1-\alpha/2)}\right\}^2$. We do not suggest using this revised constant c^* in order to gain this equivalency.

7.2.8 Logit method

The third class of methods consists of transformed methods, in which variance-stabilizing transformations, commonly used for binary data, are applied to construct the confidence interval. The point estimate is transformed to the new scale, and a Wald-type interval is constructed around the transformed point estimate, with the delta method being used to determine the transformed variance. The logit-transformed confidence interval can be constructed in the following way (Rust and Rao, 1996):

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z(1-\alpha/2)\frac{\widehat{SE}(\widehat{p})}{\widehat{p}(1-\widehat{p})} \tag{7.5}$$

The above limits are on the log odds/logit scale, and one must apply the function $\exp(\cdot)/[1 + \exp(\cdot)]$ to convert them to the standard risk scale. The limits on the logit scale

can be re-expressed using the following equivalent formulation:

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z(1-\alpha/2)\left[n_{eff}\widehat{p}(1-\widehat{p})\right]^{-1/2} \tag{7.6}$$

We refer to this as the Logit method.

To adjust for the design degrees of freedom, $z(1-\alpha/2)$ in expression 7.5 can be replaced with $t_{df_{design}}(1-\alpha/2)$. Alternatively, one can replace n_{eff} in expression 7.6 with n^*_{eff}. These intervals can be shown to be equivalent. We refer to either of these confidence interval methods as the Logit, adj. method. As before, these adjusted intervals are not equivalent if the Korn and Graubard formulation of the adjusted effective sample size ($n^*_{eff,KG}$) is used.

7.2.9 Arcsine method

The arcsine-transformed confidence interval (Hogg and Craig, 1995) can be constructed in the following way:

$$\sqrt{\widehat{p}} \pm z(1-\alpha/2)\frac{1}{2\sqrt{n_{eff}}} \tag{7.7}$$

The above limits are on the arcsine scale, and one must apply the function $[sin(\cdot)]^2$ to convert them to the standard risk scale. We refer to this as the Arcsine method.

To adjust for the design degrees of freedom, $z(1-\alpha/2)$ in expression 7.7 can be replaced with $t_{df_{design}}(1-\alpha/2)$. Alternatively, one can replace n_{eff} in expression 7.7 with n^*_{eff}. It can be shown that these approaches produce the same confidence interval limits. Again, the expressions confidence limits are not equivalent if the Korn and Graubard formulation of the adjusted effective sample size ($n^*_{eff,KG}$) is used.

7.3 Truncation and degenerate intervals

7.3.1 Truncation

In settings where it is expected that the survey will only increase the standard error relative to a simple random sample (design effect ¿ 1), such as clustered surveys, Korn and Graubard (1998) recommend using a procedure called truncation. If the observed effective sample size is greater than the actual sample size, they recommend setting n_{eff} (or $n^*_{eff,KG}$) equal to n_{act}. In other words, the observed design effect is less than 1, so we set the design effect equal to 1. Since this is equivalent to treating the data as if it resulted from a simple random sample, the logical next step would be to apply the standard IID confidence interval methods to the data, adjusting for weighting as necessary. In reality, for many formulations of the confidence intervals, after truncation, we are not left with the standard intervals. In this section we discuss these inconsistencies and recommend a revised framework to increase logical consistency. For all methods described that do not adjust for the design degrees of freedom, truncating the effective sample size n_{eff} at the actual sample size yields the equivalent IID confidence interval. For methods directly using the estimated standard error in the calculations (Wald and Logit), truncation can be achieved by checking if the estimated survey standard error is less than the simple random sample standard error, i.e., check if $\widehat{SE}(\widehat{p}) < \sqrt{\widehat{p}(1-\widehat{p}/n_{act}}$; if so, the estimated simple random sample standard error should be used for all calculations. Alternatively, since we present the Wald and Logit intervals with equivalent formulations using the effective sample size, n_{eff} can be truncated as described above.

For the methods described that do adjust for the design degrees of freedom, the truncated intervals do not always readily reduce to the standard IID interval. Consider the adjusted Wilson (Wilson, adj.) method. We have two equivalent formulations for this method, one using a t-quantile with the effective sample size, and the other using a Z-quantile with the adjusted effective sample size. Following the instructions of Korn and Graubard (1998),

we could either use the formulation with the effective sample size and truncate n_{eff}, or we could use the formulation with the adjusted effective sample size and truncate n^*_{eff}. These two approaches lead to different intervals. For example, let $n_{eff} = 60$, $n_{act} = 30$, $\widehat{p} = 0.10$, and $df_{design} = 10$, thus the observed $DEFF = 0.5$. The resulting confidence intervals are described in Table 7.1. As noted previously, the two intervals are equivalent in the absence of truncation. For this example, the comparable IID interval is (0.035, 0.256). We see that truncating the adjusted Wilson interval with the t-quantile leads to a confidence interval that is wider than the IID interval, while truncating n^*_{eff} returns the standard IID interval.

Table 7.1: Results of Truncation Example

Adjusted Intervals	**Not Truncated**	**Truncated**
Wilson, adj. t-quantile & n_{eff}	$t_{df} = 2.23$, $n_{eff} = 60$ CI = (0.042, 0.219)	$t_{df} = 2.23$, $n_{eff} = 30$ CI = (0.030, 0.283)
Wilson, adj. Z-quantile & n^*_{eff}	$Z = 1.96$, $n^*_{eff} = 46.4$ CI = (0.042, 0.219)	$Z = 1.96$, $n^*_{eff} = 30$ CI = (0.035, 0.256)

The same phenomenon occurs with all other adjusted intervals that have two equivalent forms, one with a t-quantile and the effective sample size, and the other with a Z-quantile and the adjusted effective sample size. For this reason, when applying truncation to an adjusted interval, we argue that only the adjusted effective sample size n^*_{eff} should be truncated because truncating the unadjusted effective sample size n_{eff} leads to confidence intervals wider than the equivalent IID intervals.

7.3.2 Degenerate intervals

We now discuss the behavior of each confidence interval method in the settings that can yield degenerate intervals. A degenerate interval is a confidence interval with zero width; this can occur when $\widehat{p} = 0$ or $\widehat{p} = 1$, or this can also occur when $\widehat{SE}(\widehat{p}) = 0$, which happens

if, for example, all sampled clusters in one-stage cluster sampling have the same observed prevalence of the outcome.

In settings with degenerate data, Korn and Graubard (1998) suggest that either the effective sample size, n_{eff}, or the adjusted effective sample size, $n^*_{eff,KG}$, should be set equal to n_{act}. To align this with our discussion of truncation, because the estimated variance is 0, the survey effective sample size is infinitely large, and thus a reasonable approach is to truncate it at the actual sample size. For the class of unadjusted methods, we recommend truncating the effective sample size. For the class of adjusted methods, we recommend truncating the adjusted effective sample size, n^*_{eff}. Consistent with the previous section on truncation, if the effective sample size, n_{eff}, is truncated in the adjusted formulations, we will return intervals wider than the comparable IID intervals.

Briefly we discuss the behavior of each confidence interval method in these degenerate data settings as it is important for evaluating their relative merits. We discuss only the adjusted methods because the key results are the same. We start with the setting of $\widehat{SE}(\widehat{p}) = 0$ but $\widehat{p}$ is strictly between 0 and 1. In this case, all methods are tractable, but some lead to degenerate confidence intervals. In practice, we would prefer a non-degenerate interval that reflects the sample size through the confidence interval width. The modified Wilson, Clopper-Pearson, Jeffreys, Agresti-Coull, and Arcsine methods do not lead to degenerate confidence intervals in this setting as long as the adjusted effective sample size is truncated at the actual sample size. The Wald and Logit methods lead to degenerate intervals when the standard error is zero. In this setting, we recommend using the alternative formulations for the Wald and Logit methods expressed as functions of the adjusted effective sample size; then, the adjusted effective sample size can be truncated at the actual sample size to yield non-degenerate intervals.

When the estimated proportion is equal to 0 or 1, some of the methods lead to degenerate confidence intervals, and one of the methods is not tractable. The modified Wilson, Clopper-Pearson, Jeffreys, and Agresti-Coull methods do not lead to degenerate confi-

dence intervals in this setting as long as the adjusted effective sample size is truncated at the actual sample size. The Wald method will always lead to a degenerate confidence interval, even if the adjusted effective sample size is used and properly truncated. Similarly, the Arcsine interval always produces a degenerate confidence interval; interestingly, this degenerate interval will be located slightly above 0 (or slightly below 1). The Logit method is not tractable when the estimated prevalence is equal to 0 or 1 because of the $\widehat{p}(1-\widehat{p})$ term in the denominator of the confidence interval arm. Korn and Graubard (1998) suggest substituting the truncated Clopper-Pearson interval (observed $\widehat{p}$ and n_{act}) for the Logit. In theory, any of the non-degenerate confidence interval methods could be substituted here, or an alternative method for degenerate data such as that suggested by Louis (1981) could be used.

Overall, we prefer a method that never results in a degenerate interval, which can be achieved for many of the intervals if the adjusted effective sample size is properly truncated. We provide a suggested framework for truncation in these degenerate data settings.

7.4 Simulations

7.4.1 Set-up

To assess the performance of the confidence interval methods, we performed a simple simulation study modeled after the popular 30 by 7 design used by the Expanded Programme on Immunization to estimate immunization coverage (Henderson and Sundaresan, 1982); in this design, 30 clusters are sampled with probability proportional to size and 7 children are selected within each cluster.

For our simulations, we generated 1000 primary sampling units (PSUs) with sizes drawn from a gamma distribution with shape parameter equal to 2, and scale parameter equal to 100 (mean size 200). We simulated the outcomes using a beta-binomial distribution, for

which PSU_i has prevalence p_i drawn from a $Beta(\alpha, \beta)$ distribution, and each of the secondary sampling units (SSUs) in that PSU are drawn from a Bernoulli distribution with success probability p_i. To simulate data with a particular overall prevalence p and intracluster correlation ICC, we determined that the parameters from the Beta distribution must equal the following (Ridout et al., 1999):

$$\alpha = \left[\frac{1 - ICC}{ICC}\right] p$$
$$\beta = \left[\frac{1 - ICC}{ICC}\right] (1 - p)$$

We considered prevalence values $p = 0.01, 0.02, , 0.99$, and intracluster correlations $ICC = 0.005, 0.010, 0.050, 0.100$, and 0.150, with the last resulting in a design effect of 1.9. To simulate two stage cluster sampling, we randomly sampled $n = 30$ PSUs using probability proportional to size (PPS) sampling with replacement; then, we randomly sampled $m = 7$ SSUs from each selected PSU using simple random sampling without replacement. 10,000 simulations were run for all combinations of n, m, p, and ICC.

For each simulated cluster sample, we calculated the estimated prevalence, $\widehat{p}$, which is the mean of the observed data since the data is self-weighting (Lemeshow and Robinson, 1985, eq. 1). We also estimated the variance using a standard estimator for unequal probability with-replacement two-stage sampling, simplified for the PPS setting (Lemeshow and Robinson, 1985, eq. 3) (Lohr, 2010, eq. 6.15):

$$\widehat{var}(\widehat{p}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\widehat{p}_i - \widehat{p})^2$$

where $\widehat{p}_i$ is the observed prevalence in the ith selected PSU. Then, for each sample, we calculated each of the confidence intervals described above: (1) Wald, (2) Wald, adj., (3) Wilson, (4) Wilson, adj., (5) CP, (6) CP, adj., (7) Jeffreys, (8) Jeffreys, adj., (9) AC, (10) AC, adj., (11) Logit, (12), Logit, adj., (13) Arcsine, and (14) Arcsine, adj. We summarized

the results for each method by reporting the confidence interval coverage, which is the proportion of simulations for which the estimated confidence interval contained the true value p. We also calculated the average confidence interval width.

7.4.2 Results

Simulation coverage results are shown in Figure 7.1 for $p = 0.01$ to 0.99, $n = 30$ PSUs, $m = 7$ SSUs per PSU, and $ICC = 0.15$. We do not report results for the other ICC values because the relative performance of the methods was consistent across the simulations. In general, the methods perform the worst when the ICC is large, so we report the results for the largest ICC considered.

7.4.3 Adjustment

From Figure 7.1, it is apparent that for all seven confidence interval methods, the adjusted version has superior coverage than the unadjusted version. With the exception of the adjusted Clopper-Pearson interval, all versions that are not adjusted for the design degrees of freedom have coverage below the nominal 95% level, whereas the adjusted versions have coverage near or above the 95% level. Thus, we only consider the adjusted methods henceforth.

7.4.4 Confidence interval coverage

When contrasting the individual methods, we see that the adjusted Wald interval coverage tends to fall below the nominal level when $p < 0.25$ or > 0.75. Coverage drops precipitously when $p < 0.10$ or > 0.90. The adjusted Wilson and Jeffreys intervals have coverage very close to the nominal level, though the coverage drops for p close to 0 or 1. The adjusted Agresti-Coull and Logit intervals have relatively stable coverage close to the nominal level. The adjusted Clopper-Pearson method is further above 95% than the

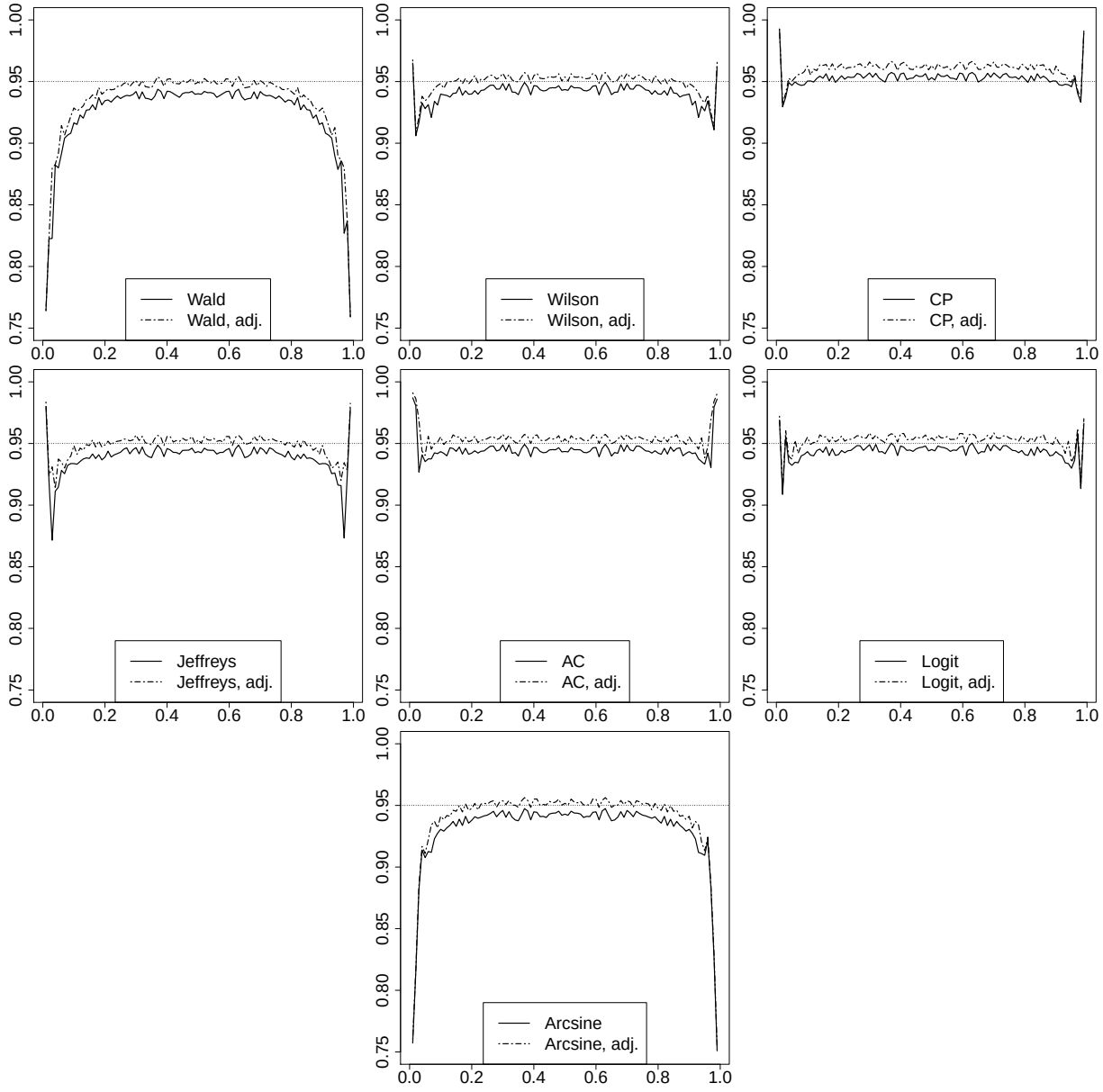

Figure 7.1: Confidence interval coverage probability versus true prevalence of outcome (0.01 to 0.99) for $n = 30$ PSUs, $m = 7$ SSUs per PSU, and $ICC = 0.15$. All methods are shown with unadjusted and adjusted intervals. (a) Wald, (b) Wilson, (c) Clopper-Pearson, (d) Jeffreys, (e) Agresti-Coull, (f) Logit, (g) Arcsine.

other methods. The adjusted Arcsine method performs well for moderate p, but coverage drops precipitously as p approaches 0 or 1.

7.4.5 Confidence interval width

To simultaneously assess confidence interval width and coverage, we plot average coverage versus average width in Figure 7.2 for each of the seven adjusted methods. The behavior of the methods changes as a function of the prevalence, so we show separate plots for prevalence values averaged over the ranges 1% to 10% in Figure 7.2(a), 10% to 25% in Figure 7.2(b), and 25% to 50% in Figure 7.2(c). The most desirable methods lie closest to the upper left-hand corner (narrowest width with the highest coverage).

For p between 1% and 10% in Figure 7.2(a), the Arcsine and Wald methods have the poorest coverage. This is likely attributable to how these intervals handle degenerate data. Degenerate data produces degenerate intervals for both the Wald and Arcsine intervals. The Logit, Jeffreys, and Wilson intervals all have coverage slightly below 95%. The Agresti-Coull and Clopper-Pearson intervals are the only intervals with coverage above 95%, but they are also the widest intervals.

For p between 10% and 25% in Figure 7.2(b), the Wald and Arcsine intervals again fail to achieve the nominal coverage level. The Jeffreys interval has coverage barely below the nominal coverage level, and it is the narrowest interval on average. The Wilson interval performs very well with the narrowest width among the methods achieving the average coverage level. The Agresti-Coull and Logit intervals are slightly wider with slightly higher coverage. The Clopper-Pearson interval is quite a bit wider than the other intervals.

For p between 25% and 50% in Figure 7.2(c), the performances of the intervals are similar to each other (note the x-axis scale), though the Clopper-Pearson interval is quite wide. The Wilson and Agresti-Coull intervals perform the best. The Logit, Jeffreys, and Arcsine

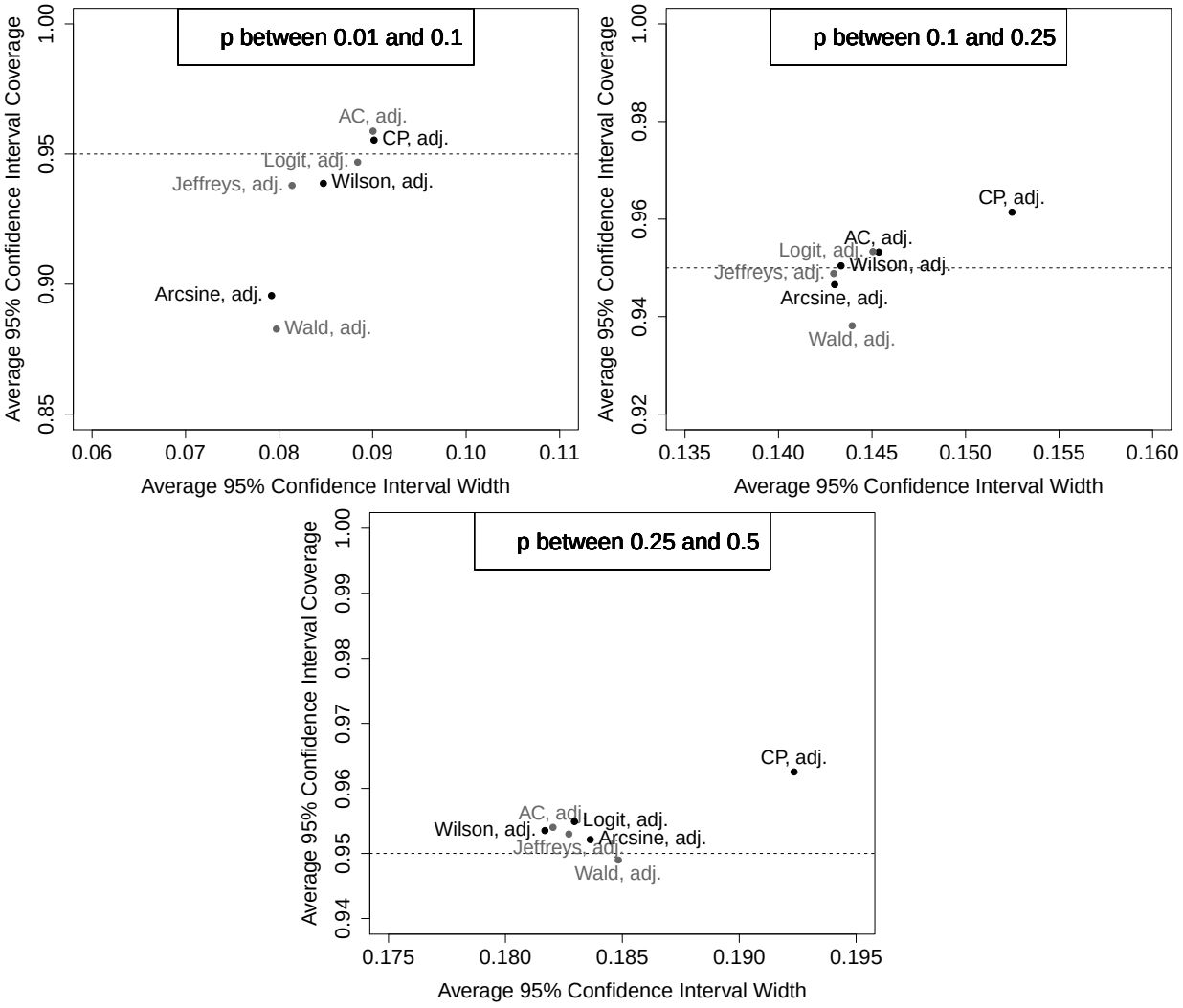

Figure 7.2: Average coverage vs. average width plot. $n = 30$, $m = 7$, $ICC = 0.15$

intervals are all slightly wider. The Wald interval performs very poorly, having the lowest coverage and still being wider than the majority of the intervals. The Clopper-Pearson interval is significantly wider than the other intervals.

7.4.6 Truncation

To assess the effect of truncation, we performed a set of simulations in which the adjusted effective sample size was truncated if it exceeded the actual sample size. Because this is most likely to occur when the ICC is small, we show results for $ICC = 0.005$. Figure 7.3 compares coverage for the adjusted Wilson with and without truncation. The effect of truncation is a moderate increase in coverage because it increases the width of confidence intervals that are narrower the comparable IID confidence interval.

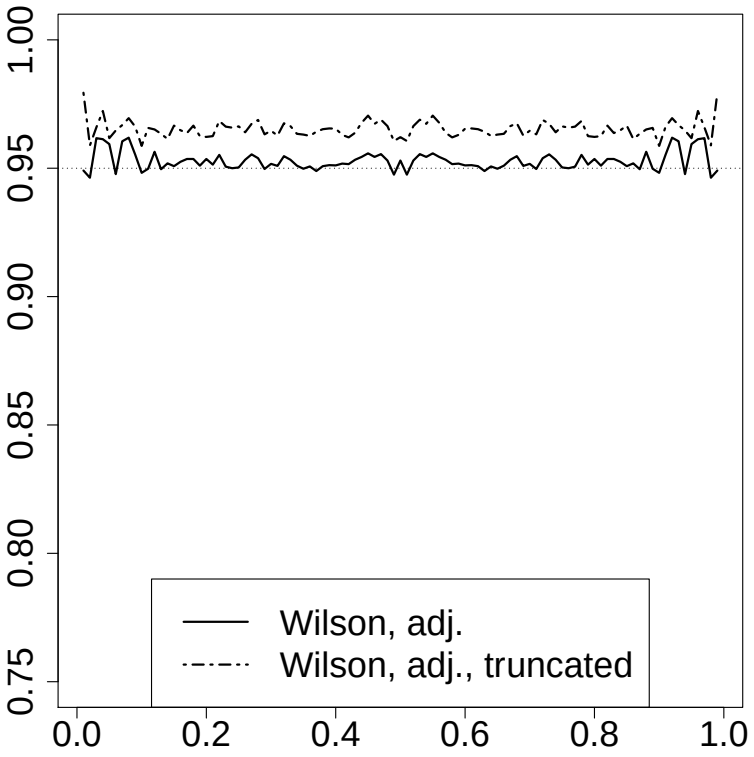

Figure 7.3: Confidence interval coverage probability versus true prevalence of outcome (0.01 to 0.99) comparing adjusted Wilson with and without truncation for $n = 30$, $m = 7$, $ICC = 0.005$

7.4.7 Non-unimodal simulation

The beta binomial prior is a rich family, but it is mostly a unimodal family, except for the uniform member. To investigate a richer prior, we generated PSU means from a mixture distribution of three betas. Each component beta distribution had an ICC of 0.005. The means we used were 5%, 15%, and 45%, with weights 20%, 40% and 40% respectively, resulting in an overall trimodal distribution. The overall mean was 25% with $ICC = 0.15$. We summarize the results of 50,000 iterations in Figure 7.4 for $n = 30$ PSUs and $m = 7$ SSUs per PSU.

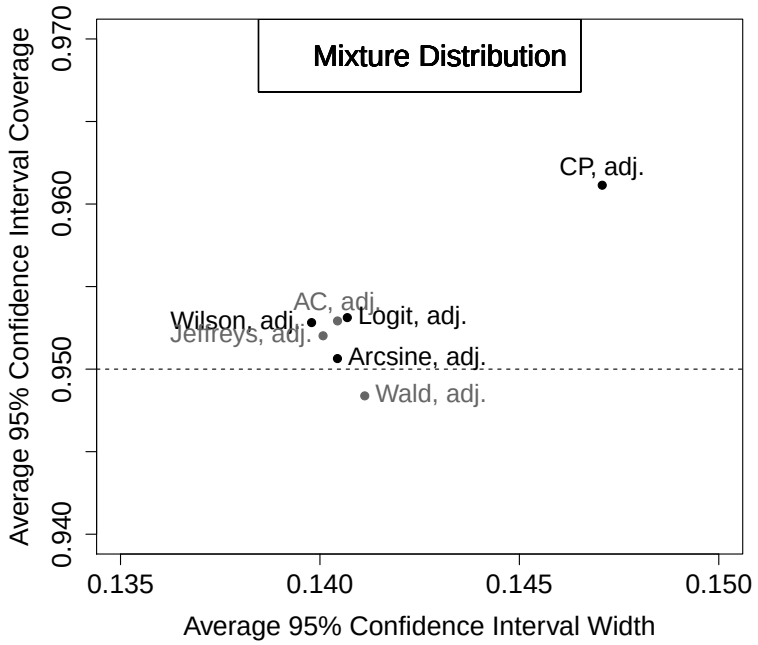

Figure 7.4: Average coverage vs. average width plot for mixture distribution

In this scenario, the Wilson interval performs the best, being both the narrowest and with high coverage. The Agresti-Coull, Logit, Jeffreys and Arcsine intervals also perform well, although the Arcsine coverage is a bit lower and the Logit and Agresti-Coull intervals are a bit wider than the Wilson. The Wald interval performs poorly, having both low coverage and high width. The Clopper-Pearson is much wider than the other intervals.

7.5 Discussion

In this paper, we describe an extensive evaluation of seven methods for constructing confidence intervals for proportions using survey data. We classified the intervals into three categories: (1) the Wald interval, (2) a class of modified methods, and (3) a class of transformed methods. To the best of our knowledge, this paper contains the first application of the modified Jeffreys interval and the first description of the modified Agresti-Coull interval. All methods described require very few parameters to calculate just an estimate of the prevalence, an estimate of the standard error, the design degrees of freedom (if the interval will be adjusted), and the actual sample size (if the interval will be truncated). Thus, these methods can be readily applied by researchers with limited statistical expertise, an aspect which is highly relevant to people analyzing the types of surveys that would emerge from a 30x7 cluster design.

For each method, we describe two formulations: one that does and one that does not adjust for the design degrees of freedom. For many of the intervals, we express the adjusted version using either a t-quantile with the effective sample size or a Z-quantile with the adjusted effective sample size. We note that these versions are only equivalent when our proposed version of the adjusted effective sample size is used. From our simulation results, we conclude that it is necessary to adjust for the design degrees of freedom in order to achieve nominal confidence interval coverage levels.

Considering the seven adjusted intervals described and the simulation results, some clear recommendations emerge. The Wald interval should be avoided entirely because it frequently fails to achieve the nominal coverage level while still being wider than the other intervals when $10\% < p < 90\%$; in addition, it can produce degenerate intervals or intervals with bounds beyond 0 or 1. Similarly, the Arcsine interval performs poorly across all simulations, partially because it can produce degenerate confidence intervals. For $25\% < p < 50\%$, the Arcsine interval is also wider than many of the other intervals. Thus, we do not recommend use of the Arcsine interval.

Among the remaining intervals, when $p < 10\%$ or $p > 90\%$, the Agresti-Coull and Clopper-Pearson intervals perform the best, although it is worth noting that the other intervals only narrowly fail to achieve the nominal coverage level. Though complete simulation results are not shown, for less extreme ICC values, these other methods do achieve 95% coverage. For all intermediate prevalence values and for the mixture distribution example, the Logit, Wilson, Jeffreys, and Agresti-Coull intervals perform well. The Wilson and Jeffreys intervals tend to be narrowest. The Logit interval tends to be wider than the other intervals. The width and performance of the Agresti-Coull interval varies across the different prevalence ranges.

We also expand upon an existing framework for truncation and the handling of 'degenerate' data described by Korn and Graubard with the goals of increasing logical consistency and making clear practical recommendations. If truncation is to be used, we demonstrate the importance of truncating the adjusted effective sample size, rather than the effective sample size, so that intervals reduce to their standard IID equivalents. In addition, we describe how each of the seven intervals performs when the standard error of the data is equal to 0 and/or the point estimate of the data is equal to 0 or 1. In these scenarios, we propose a framework in which the adjusted effective sample size is truncated at the actual sample size, thereby avoiding degenerate intervals for many of the methods.

There are limitations to our work. Our simulations studies did not address many of the other relevant factors in surveys, including stratification, disproportionate weighting, missing data, and so on. Furthermore, the simulations were constructed to resemble the particular setting of the 30x7 survey, which does not generalize to all other settings. Nonetheless, the results are consistent with the large body of literature describing methods for confidence interval construction for proportions in the IID setting, so we believe that the recommendations would not change significantly if more extensive simulations were performed.

Another limitation is that we do not apply all possible methods for confidence interval

construction. There are further adaptations of the Wilson interval described in the literature, including the use of a continuity correction (Korn and Graubard, 1999) or dropping terms that are $O_p(n^{-3/2})$ (Kott et al., 2001). Preliminary evidence in our simulations suggested that these intervals did not perform as well as the standard modified Wilson interval. Other modifications of the Wilson interval include the Andersson-Nerman interval and the Model-based Wilson interval (Kott et al., 2001). We did not investigate these intervals because they require calculations of additional quantities beyond the point estimate and the standard error. We believe that this makes the intervals less appealing to practitioners and leave their further investigation to others. The Breeze interval is another interval suggested for use in survey settings (Breeze, 1990). It is based on the Poisson distribution, so it is only appropriate for small (or large) $\widehat{p}$, and it is not expected to perform well for moderate p. There are transformations, besides the logit and arcsine functions, that have been suggested. Korn and Graubard (1999, p. 66) describe the use of a log transformation , although this interval is not guaranteed to remain within the 0 to 1 bounds. The likelihood ratio interval is obtained by inverting the likelihood ratio test $H_0 : p = p_0$ (Feng and Sitter, 2008). This interval was not included in our simulations because an iterative algorithm is required to find the bounds, which we again believe makes the interval less appealing to users. Finally, the class of replication-based methods, including the bootstrap and jackknife, comprises an important category of methods for the analysis of survey data (Rust and Rao, 1996). These intervals are interesting, but we excluded them because of the computing complexity.

Our work has important practical implications because it addresses basic questions that are not fully described in the literature. We reiterate the importance of avoiding the Wald interval, but we also provide a variety of viable alternatives that are operationally simple to calculate. Among these intervals, the modified Jeffreys and modified Agresti-Coull have not been evaluated previously. We also provide a more structured framework for handling the design degrees of freedom, data truncation, and degenerate data with the goals of inducing equivalencies and increasing logical consistency.

A. Appendices

A.1 The use of the finite population correction in survey design for national disease surveillance

A.1.1 Estimating Survey Variance in an Infinite Population

In the first stage of sampling, n PSUs are sampled without replacement using probability proportional to size (PPS) sampling into a set S defining the indices of the sampled PSUs, and m SSUs are sampled using simple random sampling without replacement from each selected PSU_i into a set S_i defining the indices of the sampled SSUs.

$$
\begin{aligned}
w_{PSU,i} &= [Pr(i \in S)]^{-1} && \text{PSU weight} \\
&= \frac{M}{nM_i} \\
w_{SSU,i} &= [Pr(j \in S_i | i \in S)]^{-1} && \text{SSU weight (equal } \forall j \in S_i) \\
&= \frac{M_i}{m} \\
w_i &= w_{PSU,i}\, w_{SSU,i} && \text{Overall weight} \\
&= \frac{M}{nM_i}\frac{M_i}{m} \\
&= \frac{M}{nm} \equiv w \\
\widehat{T} &= \sum_{i \in S} w_i \widehat{t_i} && \text{Numerator} \\
\widehat{M} &= \sum_{i \in S} w_i m && \text{Denominator} \\
&= \sum_{i \in S} \frac{M}{nm} m \\
&= M && \text{since PPS} \\
\widehat{p} &= \frac{\widehat{T}}{\widehat{M}} && \text{Prevalence estimator}
\end{aligned}
$$

The following is derived using a result from Lohr (2010, p. 229). Let $T = \sum_{i=1}^{N} M_i p_i$. We make the assumption that site size (M_i) and site prevalence (p_i) are independent.

$$
\begin{aligned}
var(\widehat{T}) &= Var_S\left(E(\widehat{T}|S)\right) + E_S\left(Var(\widehat{T}|S)\right) \\
&= Var_S\left(E\left(\sum_{i\in S} w_i\widehat{t}_i\middle|S\right)\right) + E_S\left(Var\left(\sum_{i\in S} w_i\widehat{t}_i\middle|S\right)\right) \\
&= Var_S\left(E\left(\sum_{i\in S} w_i m\hat{p}_i\middle|S\right)\right) + E_S\left(Var\left(\sum_{i\in S} w_i m\hat{p}_i\middle|S\right)\right) \\
&= Var_S\left(\sum_{i\in S} w_{PSU,i}\frac{M_i}{m}mp_i\right) + E_S\left(\sum_{i\in S} w_i^2 m^2\frac{p_i(1-p_i)}{m}\right) \\
&= \frac{1}{n}\sum_{i=1}^{N}\frac{\Pr(i\in S)}{n}\left(\frac{M_i p_i}{\Pr(i\in S)/n} - T\right)^2 \\
&\quad + E_S\left(\sum_{i=1}^{N} I(i\in S) w_{PSU,i}^2\frac{M_i^2}{m^2}m^2\frac{p_i(1-p_i)}{m}\right) \text{ by Lohr} \\
&= \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left(\frac{M_i p_i}{M_i/M} - T\right)^2 + E_S\left(\sum_{i=1}^{N} I(i\in S)\frac{M^2}{n^2 M_i^2}M_i^2\frac{p_i(1-p_i)}{m}\right) \\
&= \frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left(Mp_i - T\right)^2 + \sum_{i=1}^{N}\frac{nM_i}{M}\frac{M^2}{n^2}\frac{p_i(1-p_i)}{m} \\
&= \frac{1}{n}M^2\sum_{i=1}^{N}\frac{M_i}{M}\left(p_i - p\right)^2 + \frac{1}{nm}M\sum_{i=1}^{N}M_i p_i(1-p_i) \\
&= M^2\frac{Var_{PSU}}{n} + \frac{1}{nm}M\left[\sum_{i=1}^{N}M_i p_i - \sum_{i=1}^{N}M_i p_i^2\right] \\
&\quad \text{defining } Var_{PSU} \text{ as the between PSU variance of the } p_i \text{ terms} \\
&= M^2\frac{Var_{PSU}}{n} + \frac{1}{nm}M\left[N\ E\left[M_i p_i\right] - N\ E\left[M_i p_i^2\right]\right] \\
&= M^2\frac{Var_{PSU}}{n} + \frac{1}{nm}M\left[N\ E\left[M_i\right]E\left[p_i\right] - N\ E\left[M_i\right]E\left[p_i^2\right]\right] \\
&\quad \text{assuming } M_i \perp p_i \\
&= M^2\frac{Var_{PSU}}{n} + \frac{1}{nm}M\left[N\overline{M}p - N\overline{M}\left\{Var_{PSU} + p^2\right\}\right] \\
&= M^2\frac{Var_{PSU}}{n} + \frac{1}{nm}M\left[Mp - Mp^2\right] - \frac{M^2}{nm}Var_{PSU} \\
&= M^2\frac{Var_{PSU}}{n} + \frac{M^2}{nm}\left\{p(1-p) - Var_{PSU}\right\} \\
&= M^2\frac{Var_{PSU}}{n} + \frac{M^2}{nm}\left\{p(1-p) - ICCp(1-p)\right\}
\end{aligned}
$$

$$
\begin{aligned}
& M^2 \frac{Var_{PSU}}{n} + \frac{M^2}{nm} p(1-p)\left[1 - ICC\right] \\
var\left(\widehat{p}\right) &= var\left(\frac{\widehat{T}}{M}\right) \\
&= \left(\frac{1}{n}\right) Var_{PSU} + \left(\frac{1}{nm}\right) p(1-p)\left[1 - ICC\right]
\end{aligned}
$$

A.1.2 Estimating the Intracluster Correlation Coefficient

We can show Equation 5.1 using standard formulae for the Beta distribution and the Beta-Binomial distribution Gelman et al. (2004, pp. 576-577) and additional notation from Ridout et al. (1999). The PSU means are sampled from a Beta distribution with parameters α and β.

The grand mean is:

$$p = \frac{\alpha}{\alpha + \beta}$$

The variance of the PSU means (p_i) is:

$$Var_{PSU} \equiv \frac{\alpha\beta}{\left(\alpha + \beta\right)^2 \left(\alpha + \beta + 1\right)}$$

The intracluster correlation (ICC) is:

$$ICC = \frac{1}{\alpha + \beta + 1}$$

We now verify that Equation 5.1 is consistent with the known ICC of the Beta-binomial distribution:

$$
\begin{aligned}
ICC &= \frac{Var_{PSU}}{p(1-p)} \\
&= \frac{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}{\frac{\alpha}{(\alpha+\beta)}\frac{\beta}{(\alpha+\beta)}} \\
&= \frac{1}{\alpha+\beta+1}
\end{aligned}
$$

A.1.3 Estimating Design Effect in an Infinite Population

For the second method, we consider the design effect for a binary outcome estimated via a two-stage cluster survey in an infinite population. The denominator of the design effect is the simple random sample variance with replacement. A simplified approach to calculating the simple random sample variance is to calculate the unit variance of a beta-binomial random variable and divide this by the overall sample size (nm).

$$
\begin{aligned}
Var(\widehat{p}|\text{one unit sampled}) &= \frac{\alpha\beta(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \left(\frac{\alpha}{\alpha+\beta}\right)\left(\frac{\beta}{\alpha+\beta}\right) \\
&= p(1-p)
\end{aligned}
$$

We can then demonstrate that the design effect in an infinite population is the following:

$$
\begin{aligned}
DEFT^2 &= \frac{var(\widehat{p})}{var(\widehat{p}_{SRS,wr})} \\
&= \frac{\left(\frac{1}{n}-\frac{1}{nm}\right)Var_{PSU}+\frac{p(1-p)}{nm}}{\frac{p(1-p)}{nm}} \\
&= (m-1)\frac{Var_{PSU}}{p(1-p)}+1 \\
&= 1+(m-1)\,ICC \qquad \text{from eq.}5.1
\end{aligned}
$$

A.1.4 Comparing First Stage fpcs for PPS Sampling

Here we assess the properties of the first stage fpc of Wolter so it can be compared with the simpler first stage fpc used by Stata.

$$\begin{aligned} fpc_1 &= \left[1 - \frac{1}{n}\sum_{i \in S} Pr(i \in S)\right] \\ &= \left[1 - \frac{1}{n}\sum_{i \in S} \frac{nM_i}{M}\right] \\ &= \left[1 - \frac{1}{M}\sum_{i \in S} M_i\right] \end{aligned}$$

Interestingly, this represents the proportion of patients in the population attending clinics that were sampled, which is not the same as the proportion of patients in the population sampled.

$$\begin{aligned} E_S\left[fpc_1\right] &= E_S\left[1 - \frac{1}{M}\sum_{i \in S} M_i\right] \\ &= 1 - \frac{1}{M}\sum_{i=1}^{N} E_S\left[I(i \in S)M_i\right] \\ &= 1 - \frac{1}{M}\sum_{i=1}^{N} \frac{nM_i}{M} M_i \\ &= 1 - \frac{Nn}{M^2}\frac{1}{N}\sum_{i=1}^{N} M_i^2 \\ &= 1 - \frac{Nn}{M^2} E_S\left[M_i^2\right] \\ &= 1 - \frac{Nn}{M^2}\left\{Var_S(M_i) + E_S(M_i)^2\right\} \\ &= 1 - \frac{Nn}{M^2}\left\{Var_S(M_i) + \overline{M}^2\right\} \\ &= 1 - Nn\frac{\overline{M}^2}{M^2} - Nn\frac{Var_S(M_i)}{M^2} \end{aligned}$$

$$
\begin{aligned}
&= 1 - \frac{Nn}{N^2} - \frac{Nn}{N^2}\frac{Var_S(M_i)}{M^2/N^2} \\
&= 1 - \frac{n}{N} - \frac{n}{N}\frac{Var_S(M_i)}{\overline{M}^2} \\
&= 1 - \frac{n}{N} - \frac{n}{N}cv^2(M_i) \qquad cv^2(\cdot) \text{ coefficient of variation}
\end{aligned}
$$

We can see that the two fpc formulations will be equivalent when M_i is constant across all i. When the PSU sizes are not equal, we can see that the above formulation will result in greater variance deflation. Thus, the simpler $(1 - n/N)$ formulation is more conservative.

A.1.5 Estimating Survey Variance in a Finite Population

$$
\begin{aligned}
var(\widehat{T}) &= Var_S\left(E(\widehat{T}|S)\right) + E_S\left(Var(\widehat{T}|S)\right) \\
&= Var_S\left(E\left(\sum_{i\in S} w_i\widehat{t}_i\middle|S\right)\right) + E_S\left(Var\left(\sum_{i\in S} w_i\widehat{t}_i\middle|S\right)\right) \\
&= Var_S\left(E\left(\sum_{i\in S} w_i m\widehat{p}_i\middle|S\right)\right) + E_S\left(Var\left(\sum_{i\in S} w_i m\widehat{p}_i\middle|S\right)\right) \\
&= Var_S\left(\sum_{i\in S} w_{PSU,i}\frac{M_i}{m}mp_i\right) + E_S\left(\sum_{i\in S}\left(1-\frac{m}{M_i}\right)w_i^2m^2\frac{p_i(1-p_i)}{m}\right) \\
&\approx \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}\frac{\Pr(i\in S)}{n}\left(\frac{M_ip_i}{\Pr(i\in S)/n} - T\right)^2 \\
&\quad + E_S\left(\sum_{i=1}^{N} I(i\in S)\left(1-\frac{m}{M_i}\right)w_{PSU,i}^2\frac{M_i^2}{m^2}m^2\frac{p_i(1-p_i)}{m}\right) \\
&\quad \text{from Lohr eq. 6.8} \\
&\quad \text{Approximate first-stage FPC using } \left(1-\frac{n}{N}\right) \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left(\frac{M_ip_i}{M_i/M} - T\right)^2 \\
&\quad + E_S\left(\sum_{i=1}^{N} I(i\in S)\left(1-\frac{m}{M_i}\right)\frac{M^2}{n^2M_i^2}M_i^2\frac{p_i(1-p_i)}{m}\right) \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}\frac{M_i}{M}\left(Mp_i - T\right)^2 + \sum_{i=1}^{N}\left(1-\frac{m}{M_i}\right)\frac{nM_i}{M}\frac{M^2}{n^2}\frac{p_i(1-p_i)}{m} \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2\sum_{i=1}^{N}\frac{M_i}{M}\left(p_i - p\right)^2 + \frac{1}{nm}M\sum_{i=1}^{N}\left(1-\frac{m}{M_i}\right)M_ip_i(1-p_i)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \frac{1}{nm}M\left[\sum_{i=1}^{N} M_i p_i(1-p_i) - \sum_{i=1}^{N} m p_i(1-p_i)\right] \\
&\quad \text{defining } Var_{PSU} \text{ as the weighted between PSU variance of the } p_i \text{ terms} \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \frac{1}{nm}M\left[N\ E\left[M_i p_i(1-p_i)\right] - mN\ E\left[p_i(1-p_i)\right]\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \frac{1}{nm}M\left[N\ E\left[M_i\right]E\left[p_i(1-p_i)\right] - mN\ E\left[p_i(1-p_i)\right]\right] \\
&\quad \text{assuming } M_i \perp p_i \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \frac{M}{nm}\left[(N\ \overline{M} - mN)E\left[p_i(1-p_i)\right]\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \frac{M}{nm}\left[(M-mN)\left\{E\left[p_i\right] - E\left[p_i^2\right]\right\}\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} + \left[\left(\frac{M^2}{nm} - \frac{NM}{n}\right)\left\{p - p^2 - Var_{PSU}\right\}\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} - \left(\frac{M^2}{nm} - \frac{NM}{n}\right)Var_{PSU} + \left(\frac{M^2}{nm} - \frac{NM}{n}\right)p(1-p) \\
var\left(\widehat{p}\right) &= var\left(\frac{\widehat{T}}{M}\right) \\
&= \frac{1}{M^2}\left\{\left(1-\frac{n}{N}\right)\frac{1}{n}M^2 Var_{PSU} - \left(\frac{M^2}{nm} - \frac{NM}{n}\right)Var_{PSU} + \left(\frac{M^2}{nm} - \frac{NM}{n}\right)p(1-p)\right\} \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n}Var_{PSU} - \left(\frac{1}{nm} - \frac{N}{nM}\right)Var_{PSU} + \left(\frac{1}{nm} - \frac{N}{nM}\right)p(1-p) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)Var_{PSU} + \left(\frac{1}{nm} - \frac{1}{n\overline{M}}\right)\left[p(1-p) - ICC\ p(1-p)\right] \qquad \text{from eq.5.1} \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)Var_{PSU} + \left(\frac{1}{nm} - \frac{1}{n\overline{M}}\right)p(1-p)\left(1-ICC\right)
\end{aligned}
$$

A.1.6 Estimating Design Effect in a Finite Population

$$
\begin{aligned}
DEFT^2\left(\widehat{p}\right) &= \frac{\frac{Var_{PSU}}{n} - \frac{Var_{PSU}}{N} - \frac{Var_{PSU}}{nm} + \frac{Var_{PSU}}{n\overline{M}} + \frac{p(1-p)}{nm} - \frac{p(1-p)}{n\overline{M}}}{\frac{p(1-p)}{nm}} \\
&= m\frac{Var_{PSU}}{p(1-p)} - \frac{nm}{N}\frac{Var_{PSU}}{p(1-p)} + \frac{m}{\overline{M}}\frac{Var_{PSU}}{p(1-p)} - \frac{Var_{PSU}}{p(1-p)} + 1 - \frac{m}{\overline{M}} \\
&= mICC - \frac{nm}{N}ICC + \frac{m}{\overline{M}}ICC - ICC + 1 - \frac{m}{\overline{M}} \\
&= 1 - \frac{m}{\overline{M}} + ICC\left[m - \frac{nm}{N} + \frac{m}{\overline{M}} - 1\right]
\end{aligned}
$$

$$= \left(1 - \frac{m}{\overline{M}}\right) + ICC\left[m\left(1 - \frac{n}{N}\right) + \left(1 - \frac{m}{\overline{M}}\right)\right]$$

A.1.7 Sample Size Calculation Methods

Derivations for all sample size calculation methods. Note that these methods can return negative numbers; these should be viewed as impossible (n/a) sample size designs.

Method 1:

$$\begin{aligned}
L &= q\sqrt{\frac{p(1-p)}{k_{eff1}}} \\
k_{eff1} &= \frac{q^2 p(1-p)}{L^2} \\
k_{act1} &= nm_1 \\
k_{act1} &= k_{eff1} DEFT^2 \\
&= k_{eff1}\left[1 + ICC(m_1 - 1)\right] \\
nm_1 &= k_{eff1}\left[1 + ICC(m_1 - 1)\right] \\
m_1 &= \frac{k_{eff1}\left[1 - ICC\right]}{n - k_{eff1} ICC} \\
&= \frac{q^2 p(1-p)\left[1 - ICC\right]}{L^2 n - q^2 p(1-p) ICC}
\end{aligned}$$

Method 2:

$$\begin{aligned}
L &= q\sqrt{\left(1 - \frac{k_{eff2}}{M}\right)\frac{p(1-p)}{k_{eff2}}} \\
k_{eff2} &= \frac{q^2 p(1-p) M}{L^2 M + q^2 p(1-p)} \\
k_{act2} &= nm_2 \\
k_{act2} &= k_{eff2} DEFT^2
\end{aligned}$$

$$
\begin{aligned}
&= k_{eff2}\left[1 + ICC(m_2 - 1)\right] \\
nm_2 &= k_{eff2}\left[1 + ICC(m_2 - 1)\right] \\
m_2 &= \frac{k_{eff2}\left[1 - ICC\right]}{n - k_{eff2}ICC} \\
&= \frac{q^2 p(1-p)M\left[1 - ICC\right]}{n\left[L^2 M + q^2 p(1-p)\right] - q^2 p(1-p)M\left[ICC\right]}
\end{aligned}
$$

Method 3:

$$
\begin{aligned}
L &= q\sqrt{\frac{p(1-p)}{k_{eff3}}} \\
k_{eff3} &= \frac{q^2 p(1-p)}{L^2} \qquad \text{same as } k_{eff1} \\
k_{act3} &= nm_3 \\
k_{act3} &= k_{eff3} DEFT^2 \\
&= k_{eff3}\left[\left(1 - \frac{m_3}{\overline{M}}\right) + ICC\left(\left(1 - \frac{n}{N}\right)m_3 - \left(1 - \frac{m_3}{\overline{M}}\right)\right)\right] \\
nm_3 &= k_{eff3}\left[\left(1 - \frac{m_3}{\overline{M}}\right) + ICC\left(\left(1 - \frac{n}{N}\right)m_3 - \left(1 - \frac{m_3}{\overline{M}}\right)\right)\right] \\
m_3 &= \frac{k_{eff3}\left[1 - ICC\right]}{n + \frac{k_{eff3}}{\overline{M}} - k_{eff3}ICC\left[\left(1 - \frac{n}{N}\right) + \frac{1}{\overline{M}}\right]} \\
&= \frac{q^2 p(1-p)\overline{M}\left[1 - ICC\right]}{L^2 n\overline{M} + q^2 p(1-p) - q^2 p(1-p)\overline{M}\left[ICC\right]\left[\left(1 - \frac{n}{N}\right) + \frac{1}{\overline{M}}\right]}
\end{aligned}
$$

The corresponding predicted confidence interval half-width will be:

$$t_{n-H,0.975}\sqrt{Var(\widehat{p}_{ADJ})}$$

Note that the design degrees of freedom are $n - H$ where $n = \sum_{h=1}^{H} n_h$ is the total number of sites sampled, and H is the total number of strata. If no stratification is used prior to site sampling, $H = 1$.

A.2 Development of a viral load suppression measure adjusted for non-retention for the surveillance of acquired HIV drug resistance

A.2.1 Derivation of adjusted VLS variance estimator

This section outlines the derivation of the variance estimator for the adjusted VLS measure $\widehat{p}_{ADJ} = \widehat{V}/\widehat{S}$, with notation as defined in Section 6.3. Methodology follows Stata's section on variance estimation in the SVY manual (StataCorp, 2013), and it applies a result from Goodman (1960, eq. 7) for the variance of a product of independent random variables, referring to the independence of $\widehat{p}_{VLS,i}$ and $\widehat{p}_{RET,i}$. Recall that $\widehat{p}_{VLS,i} = \widehat{t}_i/m_i$ and $\widehat{p}_{RET,i} = \widehat{u}_i/s_i$. $\widehat{t}_i$ and $\widehat{u}_i$ are independent because they are taken from separate and unrelated samples (the former from incoming eligible patients and the latter from eligible patient records).

$$
\begin{aligned}
\widehat{var}(\widehat{V}) &= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i} - \frac{1}{n}\sum_{i'=1}^{n} w_{PSU,i'}S_{i'}\widehat{p}_{VLS,i'}\widehat{p}_{RET,i'}\right)^2 \\
&\quad + \frac{n}{N}\sum_{i=1}^{n}\widehat{var}(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i}) \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i} - \frac{\widehat{V}}{n}\right)^2 \\
&\quad + \frac{n}{N}\sum_{i=1}^{n}(w_{PSU,i}S_i)^2\Big[\widehat{p}^2_{VLS,i}\widehat{var}(\widehat{p}_{RET,i}) + \\
&\qquad + \widehat{p}^2_{RET,i}\widehat{var}(\widehat{p}_{VLS,i}) - \widehat{var}(\widehat{p}_{VLS,i})\widehat{var}(\widehat{p}_{RET,i})\Big] \\
&\quad \text{(Goodman 1960)} \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i} - \frac{\widehat{V}}{n}\right)^2 \\
&\quad + \frac{n}{N}\sum_{i=1}^{n}(w_{PSU,i}S_i)^2\left\{\widehat{p}^2_{VLS,i}\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\,(1-\widehat{p}_{RET,i})}{s_i}\right. \\
&\qquad \left. + \widehat{p}^2_{RET,i}\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\,(1-\widehat{p}_{VLS,i})}{m_i}\right.
\end{aligned}
$$

$$\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i}\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)}{s_i}\Bigg\}$$

$$\begin{aligned}
\widehat{var}(\widehat{S}) &= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i-\frac{1}{n}\sum_{j=1}^{n}w_{PSU,j}S_j\right)^2+0\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i-\frac{\widehat{S}}{n}\right)^2\\
\widehat{cov}(\widehat{V},\widehat{S}) &= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\Bigg[\left(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i}-\frac{1}{n}\sum_{j=1}^{n}w_{PSU,j}S_j\widehat{p}_{VLS,j}\widehat{p}_{RET,j}\right)\\
&\quad\times\left(w_{PSU,i}S_i-\frac{1}{n}\sum_{j=1}^{n}w_{PSU,j}S_j\right)\Bigg]+0\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\widehat{p}_{VLS,i}\widehat{p}_{RET,i}-\frac{\widehat{V}}{n}\right)\left(w_{PSU,i}S_i-\frac{\widehat{S}}{n}\right)\\
\widehat{var}(\widehat{p}_{ADJ}) &= \frac{1}{\widehat{S}^2}\left\{\widehat{var}(\widehat{V})-2\widehat{p}_{ADJ}\widehat{cov}(\widehat{V},\hat{S})+\widehat{p}_{ADJ}^2\widehat{var}(\hat{S})\right\}\\
&\cdots\\
&= \frac{1}{\widehat{S}^2}\left(1-\frac{n}{N}\right)\frac{n}{n-1}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\right)^2\left(\widehat{p}_{VLS,i}\widehat{p}_{RET,i}-\widehat{p}_{ADJ}\right)^2\\
&\quad+\frac{1}{\widehat{S}^2}\frac{n}{N}\sum_{i=1}^{n}\left(w_{PSU,i}S_i\right)^2\Bigg\{\widehat{p}_{VLS,i}^2\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)}{s_i}\\
&\qquad+\widehat{p}_{RET,i}^2\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i}\\
&\qquad+\widehat{p}_{RET,i}^2\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i}\\
&\qquad-\left(1-\frac{m_i}{M_i}\right)\frac{\widehat{p}_{VLS,i}\left(1-\widehat{p}_{VLS,i}\right)}{m_i}\left(1-\frac{s_i}{S_i}\right)\frac{\widehat{p}_{RET,i}\left(1-\widehat{p}_{RET,i}\right)}{s_i}\Bigg\}
\end{aligned}$$

A.2.2 Derivation of predicted variance in an infinite population

In this section, we outline the derivation for Equation 6.1 for the predicted variance of $\widehat{p}_{ADJ}$ for an infinite population (no finite population corrections). In order to derive this equation, we assume that sampling of PSUs is PPS (proportional to the total size of the eligible retention records, S_i). Thus, $\Pr(i \in S_I) = \frac{nS_i}{S}$ and $w_{PSU,i} = \frac{S}{nS_i}$, where $S = \sum_{i=1}^{N} S_i$. Let $V = \sum_{i=1}^{N} S_i p_{VLS,i} p_{RET,i}$. Simplifying assumptions are made to yield results

that limit the amount of prior information required to predict the variance. Among these assumptions, it is assumed that VLS and retention are independent from PSU size S_i.

First, we provide important results about estimating the between PSU variance for observed VLS, retention, and adjusted VLS. First, it can be shown that the between PSU variance for observed VLS and retention can be expressed using the following formulae (see Chapter 5).

$$Var_{PSU,VLS} = ICC_{VLS} p_{VLS}(1 - p_{VLS}) \tag{7.1}$$

$$Var_{PSU,RET} = ICC_{RET} p_{RET}(1 - p_{RET}) \tag{7.2}$$

Available data suggested no evidence of a correlation between site-specific VLS and site-specific retention (World Health Organization, 2012a, table 9). If we are willing to assume their independence, we can estimate the variance using the following formulae involving the between PSU variance of VLS, $Var_{PSU,VLS}$, and the between-PSU variance of retention, $Var_{PSU,RET}$ (Goodman, 1960, eq. 6).

$$Var_{PSU,ADJ} = p_{RET}^2 Var_{PSU,VLS} + p_{VLS}^2 Var_{PSU,RET} + Var_{PSU,VLS} Var_{PSU,RET} \tag{7.3}$$

To derive the predicted variance for adjusted VLS, we use a result from Lohr (2010, p. 229) and Goodman (1960, eq. 6) for the variance of a product of independent random variables; the assumption of independence is justified for reasons stated in Appendix A.2.1.

A.2.3 Derivation of predicted variance in a finite population

In this section, we outline the derivation for Equation 6.2 for the predicted variance of $\widehat{p}_{ADJ}$ using finite population corrections. We proceed using the same notation and many of the same results as Appendix A.2.2.

$$
\begin{aligned}
Var(\widehat{V}) &= Var_{S_I}\left(E\left[\widehat{V}|S_I\right]\right) + E_S\left[Var\left(\widehat{V}|S\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{N}\left(1-\frac{n}{N}\right)\frac{\Pr(i\in S_I)}{n}\left(\frac{S_i p_{VLS,i} p_{RET,i}}{\Pr(i\in S_I)/n} - V\right)^2 \text{ Lohr} \\
&\quad + \sum_{i=1}^{N}\left[\Pr(i\in S_I) w_{PSU,i}^2 S_i^2 \left\{ p_{RET,i}^2 Var\left(\widehat{p}_{VLS,i}\right)\right.\right. \\
&\qquad \left.\left. + p_{VLS,i}^2 Var\left(\widehat{p}_{RET,i}\right) + Var\left(\widehat{p}_{VLS,i}\right) Var\left(\widehat{p}_{RET,i}\right)\right\}\right] \qquad (7.4)
\end{aligned}
$$

(Goodman 1960)

Defining $Var_{PSU,ADJ}$ as the between PSU variance of the $p_{VLS,i}p_{RET,i}$ terms

Assuming $m,\ s$ same for all PSUs

Assuming independence of $S_i,\ p_{VLS,i},$ and $p_{RET,i}$

Assuming $S_i/S \approx M_i/M$

Replacing $Var_{PSU,VLS}$ with $ICC_{VLS}p_{VLS}(1-p_{VLS})$ by Eqs. 7.1,7.2

Assuming that $\displaystyle\sum_{i=1}^{N}\frac{1}{M_i} \approx \frac{N}{\overline{\overline{M}}}$

$$
\begin{aligned}
&\ldots \qquad (7.5) \\
&= \frac{S^2}{n}\left\{ \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{1}{\overline{\overline{M}}}\right)[1-ICC_{VLS}]\right] p_{RET}^2 p_{VLS}(1-p_{VLS}) \qquad (7.6) \right. \\
&\quad + \left[ICC_{RET} + \left(\frac{1}{s} - \frac{1}{\overline{\overline{S}}}\right)[1-ICC_{RET}]\right] p_{VLS}^2 p_{RET}(1-p_{RET}) \\
&\quad + \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{1}{\overline{\overline{M}}}\right)[1-ICC_{VLS}]\right]\left[ICC_{RET} + \left(\frac{1}{s} - \frac{1}{\overline{\overline{S}}}\right)[1-ICC_{RET}]\right] \\
&\qquad \left. \times p_{VLS}(1-p_{VLS}) p_{RET}(1-p_{RET})\right\}
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{p}_{ADJ}) &= Var\left(\frac{\widehat{V}}{S}\right) \\
&= \frac{1}{S^2} Var(\widehat{V}) \\
&\approx \frac{1}{n}\left\{ \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{1}{\overline{\overline{M}}}\right)[1-ICC_{VLS}]\right] p_{RET}^2 p_{VLS}(1-p_{VLS}) \right. \\
&\quad + \left[ICC_{RET} + \left(\frac{1}{s} - \frac{1}{\overline{\overline{S}}}\right)[1-ICC_{RET}]\right] p_{VLS}^2 p_{RET}(1-p_{RET}) \\
&\quad + \left[ICC_{VLS} + \left(\frac{1}{m} - \frac{1}{\overline{\overline{M}}}\right)[1-ICC_{VLS}]\right]\left[ICC_{RET} + \left(\frac{1}{s} - \frac{1}{\overline{\overline{S}}}\right)[1-ICC_{RET}]\right] \\
&\qquad \left. \times p_{VLS}(1-p_{VLS}) p_{RET}(1-p_{RET})\right\}
\end{aligned}
$$

If finite population corrections will be used during the survey design,

$$
\begin{aligned}
Var(\widehat{V}) \approx & \sum_{h=1}^{H} \frac{S_h^2}{n} \Bigg\{ \left[ICC_{VLS} + \left(\frac{1}{m_h} - \frac{1}{\overline{M}_h} \right) [1 - ICC_{VLS}] \right] p_{RET}^2 p_{VLS}(1 - p_{VLS}) \\
& + \left[ICC_{RET} + \left(\frac{1}{s_h} - \frac{1}{\overline{S}_h} \right) [1 - ICC_{RET}] \right] p_{VLS}^2 p_{RET}(1 - p_{RET}) \\
& + \left[ICC_{VLS} + \left(\frac{1}{m_h} - \frac{1}{\overline{M}_h} \right) [1 - ICC_{VLS}] \right] \\
& \quad \times \left[ICC_{RET} + \left(\frac{1}{s_h} - \frac{1}{\overline{S}_h} \right) [1 - ICC_{RET}] \right] \\
& \quad \times p_{VLS}(1 - p_{VLS}) p_{RET}(1 - p_{RET}) \Bigg\} \qquad (7.7)
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{p}_{ADJ}) &= Var\left(\frac{\widehat{V}}{S} \right) \\
&\approx \frac{Var(\widehat{V})}{S^2} \qquad \text{Using Eq. 7.7} \qquad (7.8)
\end{aligned}
$$

The corresponding predicted confidence interval half-width will be:

$$t_{n-H,0.975} \sqrt{Var(\widehat{p}_{ADJ})}$$

Note that the design degrees of freedom are $n - H$ where $n = \sum_{h=1}^{H} n_h$ is the total number of sites sampled, and H is the total number of strata. If no stratification is used prior to site sampling, $H = 1$.

References

Abrahams, M., J. Anderson, E. Giorgi, C. Seoighe, K. Mlisana, L. Ping, G. Athreya, F. Treurnicht, B. Keele, N. Wood, J. Salazar-Gonzalez, T. Bhattacharya, H. Chu, I. Hoffman, S. Galvin, C. Mapanje, P. Kazembe, R. Thebus, S. Fiscus, W. Wide, M. Cohen, S. Abdool Karim, B. Haynes, G. Shaw, B. Hahn, B. Korber, R. Swanstrom, and C. Williamson (2009). "Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants," *Journal of Virology*, 83, 3556–3567.

Agresti, A. and B. Coull (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions," *The American Statistician*, 52, 119–126.

Allam, O., S. Samarani, and A. Ahmad (2011). "Hammering out HIV-1 incidence with Hamming distance," *AIDS*, 25, 2047–2048.

Andersson, E., W. Shao, I. Bontell, F. Cham, D. Cuong, A. Wondwossen, L. Morris, G. Hunt, A. Sönnerberg, S. Bertagnolio, F. Maldarelli, and M. Jordan (2013). "Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys," *Infection, Genetics and Evolution*, 18, 125–131.

Bar, K., H. Li, A. Chamberland, C. Tremblay, J. Routy, T. Grayson, C. Sun, S. Wang, G. Learn, C. Morgan, J. Schumacher, B. Haynes, B. Keele, B. Hahn, and G. Shaw (2010). "Wide variation in the multiplicity of HIV-1 infection among injection drug users," *Journal of Virology*, 84, 6241–6247.

Barin, F., L. Meyer, R. Lancar, C. Deveau, M. Gharib, A. Laporte, J. Desenclos, and D. Costagliola (2005). "Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use of dried serum spots," *Journal of Clinical Microbiology*, 43, 4441–4447.

Barnighäusen, T., T. McWalter, Z. Rosner, M. Newell, and A. Welte (2010). "HIV incidence estimating using the BED capture enzyme immunoassay," *Epidemiology*, 21, 685–697.

Barnighäusen, T., C. Wallrauch, A. Welte, T. McWalter, N. Mbizana, J. Viljoen, N. Graham, F. Tanswer, A. Puren, and M. Newell (2008). "HIV incidence in rural South Africa: Comparison of estimates from longitudinal surveillance and cross-sectional cBED assay testing," *PLoS ONE*, 3, e3640.

Bonhoeffer, S., E. Holmes, and M. Nowak (1995). "Causes of HIV diversity," *Nature*, 376, 125.

Breeze, E. (1990). *General Household Survey: Report on Sampling Error*, London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).

Brookmeyer, R. (1991). "Reconstruction and future trends of the AIDS epidemic in the United States," *Science*, 253, 37–42.

Brookmeyer, R. (1999). "Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence," *Biometrics*, 55, 608–612.

Brookmeyer, R. (2009a). "Response to correspondence on 'should biomarker estimates of hiv incidence be adjusted?'," *AIDS*, 23, 2066–2068.

Brookmeyer, R. (2009b). "Should biomarker estimates of HIV incidence be adjusted?" *AIDS*, 23, 485–491.

Brookmeyer, R., J. Konikoff, O. Laeyendecker, and S. Eshleman (2013a). "Estimation of HIV incidence using multiple biomarkers," *American Journal of Epidemiology*, 177, 264–272.

Brookmeyer, R., O. Laeyendecker, D. Donnell, and S. Eshleman (2013b). "Cross-sectional HIV incidence estimation in HIV prevention research," *Journal of Acquired Immune Deficiency Syndromes*, 63, S233–S239.

Brookmeyer, R. and T. Quinn (1995). "Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests," *American Journal of Epidemiology*, 141, 166–172.

Brookmeyer, R., T. Quinn, M. Shepherd, S. Mehendale, J. Rodrigues, and R. Bollinger (1995). "The AIDS epidemic in India: A new method for estimating current human immunodeficiency virus (HIV) incidence rates," *American Journal of Epidemiology*, 142, 709–713.

Brown, L., T. Cai, and A. DasGupta (2001). "Interval estimation for a binomial proportion," *Statistical Science*, 16, 101–133.

Busch, M., C. Pilcher, T. Mastro, J. Kaldor, G. Vercauteren, W. Rodriguez, C. Rousseau, T. Rehle, A. Welte, M. Averill, J. Garcia Calleja, and the WHO Working Group on HIV Incidence Assays (2010). "Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation," *AIDS*, 24, 2763–2771.

Chawla, A., G. Murphy, C. Donnelly, C. Booth, M. Johnson, J. Parry, A. Phillips, and A. Geretti (2007). "Human immunodeficiency virus (HIV) antibody avidity testing to identify recent infection in newly diagnosed HIV type 1 (HIV-1) seropositive persons infected with diverse HIV-1 subtypes," *Journal of Clinical Microbiology*, 45, 415–420.

Clopper, C. and E. Pearson (1934). "The use of confidence or fiducial limits illustrated in the case of the binominal," *Biometrika*, 26, 404–413.

Coffin, J. (1995). "HIV population dynamics in vivo: Implications for genetic variation, pathogenesis, and therapy," *Science*, 267, 483–489.

Cousins, M., J. Konikoff, O. Laeyendecker, C. Celum, S. Buchbinder, G. Seage, G. Kirk, R. Moore, S. Mehta, J. Margolick, J. Brown, K. Mayer, J. Koblin, D. Wheeler, J. Justman, S. Hodder, T. Quinn, R. Brookmeyer, and S. Eshleman (2014). "HIV diversity as a biomarker for HIV incidence estimation: Including a high-resolution melting diversity assay in a multiassay algorithm," *Journal of Clinical Microbiology*, 52, 115–121.

Cousins, M., O. Laeyendecker, G. Beauchamp, R. Brookmeyer, W. Towler, S. Hudelson, L. Khaki, B. Koblin, M. Chesney, R. Moore, G. Kelen, T. Coates, C. Celum, S. Buchbinder, G. Seage, T. Quinn, D. Donnell, and S. Eshleman (2011). "Use of a high resolution melting (HRM) assay to compare gag, pol, and env diversity in adults with different stages of HIV infection," *PLoS ONE*, 6, e27211.

Curtin, L., D. Kruszon-Moran, M. Carroll, and X. Li (2006). "Estimation and analytic issues for rare events in NHANES," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2893–2903.

Delwart, E., M. Magierowska, M. Royz, B. Foley, L. Peddada, R. Smith, C. Heldebrant, A. Conrad, and M. Busch (2002). "Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection," *AIDS*, 16, 189–195.

Delwart, E., H. Pan, H. Sheppard, D. Wolpert, A. Neumann, B. Korber, and J. Mullins (1997). "Slower evolution of human immunodeficiency virus type 1 quasispecies during progression to AIDS," *Journal of Virology*, 71, 7498–7508.

Delwart, E., H. Sheppard, B. Walker, J. Goudsmit, and J. Mullins (1994). "Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays," *Journal of Virology*, 68, 6672–6683.

Delwart, E., E. Shpaer, J. Louwagie, F. McCutchan, M. Grez, H. Rubsamen-Waigmann, and J. Mullins (1993). "Genetic relationships determined by a DNA heteroduplex mobility assay: Analysis of HIV-1 env genes," *Science*, 262, 1257–1261.

Dobbs, T., S. Kennedy, C. Pau, J. McDougal, and B. Parekh (2004). "Performance characteristics of the immunoglobulin G-capture BED-enzyme immunoassay, an assay to detect recent human immunodeficiency virus type 1 seroconversion," *Journal of Clinical Microbiology*, 42, 2623–2628.

Donner, A. and J. Koval (1980). "The estimation of the intraclass correlation in the analysis of family data," *Biometrics*, 36, 19–25.

Family Health International (2000). "Behavior Surveillance Surveys (BSS): Guidelines for repeated behavioral surveys in populations at risk of HIV," Arlington, VA.

Family Health International (2009). "Proceedings of the meeting on the development of assays to estimate HIV incidence," Chapel Hill, North Carolina, USA.

Feng, C. and R. Sitter (2008). "Confidence intervals for proportions and quantiles under two-stage sampling designs: An empirical study," in *Proceedings of the Survey Methods Section*, American Statistical Association.

Feng, X. (2006). "On confidence intervals for proportions with focus on the U.S. National Health and Nutrition Examination Surveys," Simon Fraser University, Department of Statistics and Actuarial Science, m.S. project.

Fiebig, E., D. Wright, B. Rawal, P. Garrett, R. Schumacher, L. Peddada, C. Heldebrant, R. Smith, A. Conrad, S. Kleinman, and M. Busch (2003). "Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection," *AIDS*, 17, 1871–1879.

Frost, S., T. Wrin, D. Smith, S. Kosakovsky Pond, Y. Liu, E. Paxinos, C. Chappey, J. Galovich, J. Beauchaine, C. Petropoulos, S. Little, and D. Richman (2005). "Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection," *Proceedings of the National Academy of Sciences*, 102, 18514–18519.

Gail, M. and R. Brookmeyer (1988). "Methods for projecting course of acquired immunodeficiency syndrome epidemic," *Journal of the National Cancer Institute*, 80, 900–911.

Ganeshan, S., R. Dickover, B. Korber, Y. Bryson, and S. Wolinsky (1997). "Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease," *Journal of Virology*, 71, 663–677.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC, 2 edition.

Ghys, P., T. Brown, N. Grassly, G. Garnett, K. Stanecki, J. Stover, and N. Walker (2004). "The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics," *Sexually Transmitted Infections*, 80, i5–i9.

Goodman, L. (1960). "On the exact variance of products," *Journal of the American Statistical Association*, 55, 708–713.

Gottlieb, G., D. Nickle, M. Jensen, K. Wong, J. Grobler, F. Li, S. Liu, C. Rademeyer, G. Learn, S. Abdool Karim, C. Williamson, L. Corey, J. Margolick, and J. Mullins (2004). "Dual HIV-1 infection associated with rapid disease progression," *Lancet*, 363, 619–622.

Gray, A., S. Haslett, and G. Kuzmicich (2004). "Confidence intervals for proportions estimated from complex sample designs," *Journal of Official Statistics*, 20, 705–723.

Grobler, J., C. Gray, C. Rademeyer, C. Seoighe, G. Ramjee, S. Abdool Karim, L. Morris, and C. Williamson (2004). "Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers," *The Journal of Infectious Diseases*, 190, 1355–1359.

Guy, R., A. Breschkin, C. Keenan, M. Catton, A. Enriquez, and M. Hellard (2005). "Improving HIV surveillance in Victoria: The role of the "detuned" enzyme immunoassay," *Journal of Acquired Immune Deficiency Syndromes*, 38, 495–499.

Haaland, R., P. Hawkins, J. Salazar-Gonzalez, A. Johnson, A. Tichacek, E. Karita, O. Manigart, J. Mulenga, B. Keele, G. Shaw, B. Hahn, S. Allen, C. Derdeyn, and E. Hunter (2009). "Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1," *PLoS Pathogens*, 5, e1000274.

Hall, H., R. Song, P. Rhodes, J. Prejean, Q. An, L. Lee, J. Karon, R. Brookmeyer, E. Kaplan, M. McKenna, R. Janssen, and the HIV Incidence Surveillance Group (2008). "Estimation of HIV incidence in the united states," *Journal of the American Medical Association*, 300, 520–529.

Hallett, T., P. Ghys, T. Barnighäusen, P. Yan, and G. Garnett (2009). "Errors in 'BED'-derived estimates of HIV incidence will vary by place, time and age," *PLoS ONE*, 4, e5720.

Hallett, T., B. Zaba, J. Todd, B. Lopman, W. Mwita, S. Biraro, S. Gregson, and J. Boerma (2008). "Estimating incidence from prevalence in generalised HIV epidemics: Methods and validation," *PLoS Medicine*, 5, 611–622.

Hargrove, J. (2009). "BED estimates of HIV incidence must be adjusted," *AIDS*, 23, 2061–2062.

Hargrove, J., J. Humphrey, K. Mutasa, B. Parekh, J. McDougal, R. Ntozini, H. Chidawanyika, L. Moulton, B. Ward, K. Nathoo, P. Iliff, and E. Kopp (2008). "Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay," *AIDS*, 22, 511–518.

Hecht, F., M. Busch, B. Rawal, M. Webb, E. Rosenberg, M. Swanson, M. Chesney, J. Anderson, J. Levy, and J. Kahn (2002). "Use of laboratory tests and clinical symptoms for identification of primary HIV infection," *AIDS*, 16, 1119–1129.

Henderson, R. and T. Sundaresan (1982). "Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method," *Bulletin of the World Health Organization*, 60, 253–260.

Ho, D., N. AU, A. Perelson, W. Chen, J. Leonard, and M. Markowitz (1995). "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection," *Nature*, 373, 123–126.

Hogg, R. and A. Craig, eds. (1995). *Introduction into Mathematical Statistics*, Englewood Cliffs, NJ: Prentice Hall.

Holmes, E., L. Zhang, P. Simmonds, C. Ludlam, and A. Leigh Brown (1992). "Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient," *Proceedings of the National Academy of Sciences*, 89, 4835–4839.

Hong, L., F. Ketema, A. Sill, K. Kreisel, F. Cleghorn, and N. Constantine (2007). "A simple and inexpensive particle agglutination test to distinguish recent from established HIV-1 infection," *International Journal of Infectious Diseases*, 11, 459–465.

Janssen, R., G. Satten, S. Stramer, B. Rawal, T. O'Brien, B. Weiblen, F. Hecht, N. Jack, F. Cleghorn, J. Kahn, M. Chesney, and M. Busch (1998). "New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes," *Journal of the American Medical Association*, 280, 42–48.

Ji, J. and L. Loeb (1994). "Fidelity of HIV-1 reverse transcriptase copying a hypervariable region of the HIV-1 env gene," *Virology*, 199, 323–330.

Jordan, M., D. Bennett, S. Bertagnolio, C. Gilks, and D. Sutherland (2008). "World Health Organization surveys to monitor HIV drug resistance prevention and associated factors in sentinel antiretroviral treatment sites," *Antiviral Therapy*, 13, 15–23.

Kalton, G., J. Brick, and T. L$\widehat{e}$ (2005). "Chapter VI: Estimating components of design effects for use in sample design," in United Nations Department of Economic and Social Affairs, Statistics Division, ed., *Household Sample Surveys in Developing and Transition Countries*, New York: United Nations, 95–122.

Karita, E., M. Price, E. Hunter, E. Chomba, S. Allen, L. Fei, A. Kamali, E. Sanders, O. Anzala, M. Katende, N. Ketter, and the IAVI Collaborative Seroprevalence & Incidence Study Team (2007). "Investigating the utility of the HIV-1 BED capture enzyme immunoassay using cross-sectional and longitudinal seroconverter specimens from Africa," *AIDS*, 21, 403–408.

Karon, J., R. Song, R. Brookmeyer, E. Kaplan, and H. Hall (2008). "Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results," *Statistics in Medicine*, 27, 4617–4633.

Kaufman, L. and P. Rousseeuw (1990). *Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley.

Kearney, M., F. Maldarelli, W. Shao, J. Margolick, E. Daar, J. Mellors, V. Rao, J. Coffin, and S. Palmer (2009). "Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals," *Journal of Virology*, 83, 2715–2727.

Keele, B. (2010). "Identifying and characterizing recently transmitted viruses," *Current Opinion in HIV and AIDS*, 5, 327–334.

Keele, B., E. Giorgi, J. Salazar-Gonzalez, J. Decker, K. Pham, M. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, X. Wei, C. Jiang, J. Kirchherr, F. Gao, J. Anderson, L. Ping, R. Swanstrom, G. Tomaras, W. Blattner, P. Goepfert, J. Kilby, M. Saag, E. Delwart, M. Busch, M. Cohen, D. Montefiori, B. Haynes, B. Gaschen, G. Athreya, H. Lee, N. Wood, C. Seoighe, A. Perelson, T. Bhattacharya, B. Korber, B. Hahn, and G. Shaw (2008). "Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection," *Proceedings of the National Academy of Sciences*, 105, 7552–7557.

Killian, M., P. Norris, B. Rawal, M. Lebedeva, F. Hecht, J. Levy, and M. Busch (2006). "The effects of early antiretroviral therapy and its discontinuation on the HIV-1 specific antibody response," *AIDS Research and Human Retroviruses*, 22, 640–647.

Kim, A., J. McDougal, J. Hargrove, T. Rehle, V. Pillay-Van Wyk, A. Puren, A. Ekra, M. Borget-Alloue, C. Adje-Toure, A. Sheikh, L. Odawo, L. Marum, and B. Parekh (2010). "Evaluating the BED capture enzyme immunoassay to estimate HIV incidence among adults in three countries in Sub-Saharan Africa," *AIDS Research and Human Retroviruses*, 26, 1051–1061.

Kish, L. (1965). *Survey Sampling*, New York: Wiley.

Kish, L. (1995). "Methods for design effects," *Journal of Official Statistics*, 11, 55–77.

Korber, B., K. Kunstman, B. Patterson, M. Furtado, M. McEvilly, R. Levy, and S. Wolinsky (1994). "Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences," *Journal of Virology*, 68, 7467–7481.

Korn, E. and B. Graubard (1998). "Confidence interval for proportions with small expected number of positive counts estimated from survey data," *Survey Methodology*, 24, 193–201.

Korn, E. and B. Graubard, eds. (1999). *Analysis of Health Surveys*, New York: John Wiley & Sons.

Kothe, D., R. Byers, S. Caudill, G. Satten, R. Janssen, W. Hannon, and J. Mei (2003). "Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence," *Journal of Acquired Immune Deficiency Syndromes*, 33, 625–634.

Kott, P., P. Andersson, and O. Nerman (2001). "Two-sided coverage intervals for small proportions based on survey data," in *Proceedings from FCSM (Federal Committee on Statistical Methodology) Research Conference*.

Kott, P. and D. Carr (1997). "Developing an estimation strategy for a pesticide data program," *Journal of Official Statistics*, 13, 367–383.

Kouyos, R., V. von Wyl, S. Yerly, J. Böni, P. Rieder, B. Joos, P. Taffè, C. Shah, P. Bürgisser, T. Klimkait, R. Weber, B. Hirschel, M. Cavassini, A. Rauch, M. Battegay, P. Vernazza, E. Bernasconi, B. Ledergerber, S. Bonhoeffer, G. HF, and the Swiss HIV Cohort Study (2011). "Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection," *Clinical Infectious Diseases*, 52, 532–539.

Laeyendecker, O., R. Rothman, C. Henson, B. Horne, K. Ketlogetswe, C. Kraus, J. Shahan, G. Kelen, and T. Quinn (2008). "The effect of viral suppression on cross sectional incidence testing in the Johns Hopkins Hospital Emergency Department," *Journal of Acquired Immune Deficiency Syndromes*, 48, 211–215.

Le Vu, S., L. Meyer, F. Cazein, J. Pillonel, C. Semaille, F. Barin, and J. Desenclos (2009). "Performance of an immunoassay at detecting recent infection among reported HIV diagnoses," *AIDS*, 23, 1773–1779.

Lee, H., E. Giorgi, B. Keele, B. Gaschen, G. Athreya, J. Salazar-Gonzalez, K. Pham, P. Goepfert, J. Kilby, M. Saag, E. Delwart, M. Busch, B. Hahn, G. Shaw, B. Korber, T. Bhattacharya, and A. Perelson (2009). "Modeling sequence evolution in acute HIV-1 infection," *Journal of Theoretical Biology*, 261, 341–360.

Lee, H., A. Perelson, S. Park, and T. Leitner (2008). "Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression," *PLoS Computational Biology*, 4, e1000240.

Leitner, T. and J. Albert (1999). "The molecular clock of HIV-1 unveiled through analysis of a known transmission history," *Proceedings of the National Academy of Sciences*, 96, 10752–10757.

Lemeshow, S. and D. Robinson (1985). "Surveys to measure program coverage and impact: A review of the methodology used by the Expanded Programme on Immunization," *World Health Statistics Quarterly*, 38, 65–75.

Lohr, S. (2010). *Sampling: Design and Analysis*, Boston: Brooks/Cole, 2 edition.

Long, E., J. Martin, J. Kreiss, S. Rainwater, L. Lavreys, D. Jackson, J. Rakwar, K. Mandaliya, and J. Overbaugh (2000). "Gender differences in HIV-1 diversity at time of infection," *Nature Medicine*, 6, 71–75.

Louis, T. (1981). "Confidence intervals for a binomial parameter after observing no successes," *The American Statistician*, 35, 154.

Marinda, E., J. Hargrove, W. Preiser, H. Slabbert, G. van Zyl, J. Levin, L. Moulton, A. Welte, and J. Humphrey (2010). "Significantly diminished long-term specificity of the BED capture enzyme immunoassay among patients with HIV-1 with very low CD4 counts and those on antiretroviral therapy," *Journal of Acquired Immune Deficiency Syndromes*, 53, 496–499.

Marston, M., K. Harriss, and E. Slaymaker (2008). "Non-response bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys: a study of nine national surveys," *Sexually Transmitted Infections*, 84, i71–i77.

Martin, D., P. Lemey, M. Lott, V. Moulton, D. Posada, and P. Lefeuvre (2010). "RDP3: a flexible and fast computer program for analyzing recombination," *Bioinformatics*, 26, 2462–2463.

McDougal, J. (2009). "BED estimates of HIV incidence must be adjusted," *AIDS*, 23, 2064–2065.

McDougal, J., B. Parekh, M. Peterson, B. Branson, T. Dobbs, M. Ackers, and M. Gurwith (2006). "Comparison of HIV type 1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay," *AIDS Research and Human Retroviruses*, 22, 945–952.

McMahon, J., J. Elliott, S. Bertagnolio, R. Kubiak, and M. Jordan (2013). "Viral suppression after 12 months of antiretroviral therapy in low- and middle-income countries: a systematic review," *Bulletin of the World Health Organization*, 91, 246–251.

McWalter, T. and A. Welte (2008). "On the estimation of the proportion of true recent infections using the BED capture enzyme immunoassay," .

McWalter, T. and A. Welte (2009). "A comparison of biomarker based incidence estimators," *PloS ONE*, 4, e7368.

Moyo, S., T. LeCuyer, R. Wang, S. Gaseitsiwe, J. Weng, R. Musonda, H. Bussmann, M. Mine, S. Engelbrecht, J. Makhema, R. Marlink, M. Baum, V. Novitsky, and M. Essex (2014). "Evaluation of the false recent classification rates of multiassay algorithms in estimating HIV type 1 subtype C incidence," *AIDS Research and Human Retroviruses*, 30, 29–36.

Novitsky, V., S. Lagakos, M. Herzig, C. Bonney, L. Kebaabetswe, R. Rossenkhan, D. Nkwe, L. Margolin, R. Musonda, S. Moyo, E. Woldegabriel, E. van Widenfelt, J. Makhema, and M. Essex (2009). "Evolution of proviral gp120 over the first year of HIV-1 subtype C infection," *Virology*, 383, 47–59.

Novitsky, V., R. Wang, L. Margolin, J. Baca, R. Rossenkhan, S. Moyo, E. van Widenfelt, and M. Essex (2011). "Transmission of single and multiple viral variants in primary HIV-1 subtype C infection," *PLoS ONE*, 6, e16714.

Obuchowski, N. (1997). "Nonparametric analysis of clustered ROC curve data," *Biometrics*, 53, 567–578.

Office of Global AIDS Coordinator (2006). "Interim recommendations for the use of the BED capture enzyme immunoassay for incidence estimation and surveillance - statement from the surveillance and survey and the laboratory working groups to the office of the global AIDS coordinator," Washington, DC, USA.

Overbaugh, J. and C. Bangham (2001). "Selection forces and constraints of retroviral sequence variation," *Science*, 292, 1106–1109.

Parekh, B., D. Hanson, H. J, B. Branson, T. Green, T. Dobbs, N. Constantine, J. Overbaugh, and J. McDougal (2011). "Determination of mean recency period for estimation of HIV type 1 incidence with the BED-capture EIA in persons infected with diverse subtypes," *AIDS Research and Human Retroviruses*, 27, 265–273.

Parekh, B., D. Hu, S. Vanichseni, G. Satten, D. Candal, N. Young, D. Kitayaporn, L. Srisuwanvilai, S. Rakhtam, R. Janssen, K. Choopanya, and T. Mastro (2001). "Evaluation of a sensitive/less-sensitive testing algorithm using the 3A11-LS assay for detecting recent HIV seroconversion among HIV individuals with HIV-1 subtype B or E infection in Thailand," *AIDS Research and Human Retroviruse*, 17, 453–458.

Parekh, B., M. Kennedy, T. Dobbs, C. Pau, R. Byers, T. Green, D. Hu, S. Vanichseni, N. Young, K. Choopanya, T. Mastro, and J. McDougal (2002). "Quantitative detection of increasing HIV type 1 antibodies after seroconversion: A simple assay for detecting recent HIV infection and estimating incidence," *AIDS Research and Human Retroviruses*, 18, 295–307.

Parekh, B. and J. McDougal (2001). "New approaches for detecting recent HIV-1 infection," *AIDS Reviews*, 3, 183–193.

Park, I. and H. Lee (2004). "Design effects for the weighted mean and total estimators under complex survey sampling," *Statistics Canada*, 30, 183–193.

Park, S., T. Love, J. Nelson, S. Thurston, A. Perelson, and H. Lee (2011). "Designing a genome-based HIV incidence assay with high sensitivity and specificity," *AIDS*, 25, F13–F19.

Piantadosi, A., B. Chohan, D. Panteleeff, J. Baeten, K. Mandaliya, J. Ndinya-Achola, and J. Overbaugh (2009). "HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response," *AIDS*, 23, 579–587.

Poss, M., A. Rodrigo, J. Gosink, G. Learn, D. Panteleeff, H. Martin, J. Bwayo, J. Kreiss, and J. Overbaugh (1998). "Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1," *Journal of Virology*, 72, 8240–8251.

Quinn, T., R. Brookmeyer, R. Kline, M. Shepherd, R. Paranjape, S. Mehendale, D. Gadkari, and R. Bollinger (2000). "Feasibility of pooling sera for HIV-1 viral RNA to diagnose acute primary HIV-1 infection and estimate HIV incidence," *AIDS*, 14, 2751–2757.

R Core Team (2012). "R: A language and environment for statistical computing," Vienna, Austria, http://www.Rproject.org.

Rawal, B., A. Degula, L. Lebedeva, R. Janssen, F. Hecht, H. Sheppard, and M. Busch (2003). "Development of a new less-sensitive enzyme immunoassay for detection of early HIV-1 infection," *Journal of Acquired Immune Deficiency Syndromes*, 33, 349–355.

Rehle, T., T. Hallett, O. Shisana, V. Pillay-van Wyk, K. Zuma, H. Carrara, and S. Jooste (2010). "A decline in new HIV infections in South Africa: Estimating HIV incidence from three national HIV surveys in 2002, 2005, and 2008," *PLoS ONE*, 5, e11094.

Ridout, M., C. Demetrio, and D. Firth (1999). "Estimating intraclass correlation for binary data," *Biometrics*, 55, 137–148.

Ritola, K., C. Pilcher, S. Fiscus, N. Hoffman, J. Nelson, K. Kitrinos, C. Hicks, J. Eron, and R. Swanstrom (2004). "Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1," *Journal of Virology*, 78, 11208–11218.

Rose, P. and B. Korber (2000). "Detecting hypermutations in viral sequences with an emphasis on G $--$ $>$ A hypermutation," *Bioinformatics*, 16, 400–401.

Rothman, K., S. Greenland, and T. Lash, eds. (2008). *Modern Epidemiology*, Philadelphia: Lippincott, Williams & Wilkins, 3 edition.

Rust, K. and V. Hsu (2007). "Confidence intervals for statistics for categorical variables from complex samples," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3134–3141.

Rust, K. and J. Rao (1996). "Variance estimation techniques for complex surveys using replication techniques," *Statistical Methods in Medical Research*, 5, 283–310.

Rutherford, G., S. Schwarcz, and W. McFarland (2000). "Surveillance for incident HIV infection: New technology and new opportunities," *Journal of Acquired Immune Deficiency Syndromes*, 25, S115–S119.

Sagar, M., E. Kirkegaard, E. Long, C. Celum, S. Buchbinder, E. Daar, and J. Overbaugh (2004). "Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes," *Journal of Virology*, 78, 7279–7283.

Sakarovitch, C., A. Alioum, D. Ekouevi, P. Msellati, V. Leroy, and F. Dabis (2007). "Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics," *Statistics in Medicine*, 26, 320–335.

Salazar-Gonzalez, J., E. Bailes, K. Pham, M. Salazar, M. Guffey, B. Keele, C. Derdeyn, P. Farmer, E. Hunter, S. Allen, O. Manigart, J. Mulenga, J. Anderson, R. Swanstrom, B. Haynes, G. Athreya, B. Korber, P. Sharp, G. Shaw, and B. Hahn (2008). "Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing," *Journal of Virology*, 82, 3952–3970.

Särndal, C., B. Swensson, and J. Wretman (2013). *Model Assisted Survey Sampling*, New York: Springer.

Sato, K., S. Miyazaki, and M. Ohya (1998). "Analysis of HIV by entropy evolution rate," *Amino Acids*, 14, 343–352.

Satten, G. and I. Longini (1994). "Estimation of incidence of HIV infection using cross-sectional marker surveys," *Biometrics*, 50, 675–688.

Schüpbach, J., M. Gebhardt, Z. Tomasik, C. Niederhauser, S. Yerly, P. Bürgisser, L. Matter, M. Gorgievski, R. Dubs, D. Schultze, I. Steffen, C. Andreutti, G. Martinetti, B. Güntert, R. Staub, S. Daneel, and P. Vernazza (2007). "Assessment of recent HIV-1 infection by a line immunoassay for HIV-1/2 confirmation," *PLoS Medicine*, 4, 1921–1930.

Searle, L. (1971). *Linear Models*, New York: Wiley.

Selleri, M., N. Orchi, M. Zaniratti, R. Bellagamba, A. Corpolongo, C. Angeletti, G. Ippolito, M. Capobianchi, and E. Girardi (2007). "Effective highly active antiretroviral therapy in patients with primary HIV-1 infection prevents the evolution of the avidity of HIV-1-specific antibodies," *Journal of Acquired Immune Deficiency Syndromes*, 46, 145–150.

Shankarappa, R., P. Gupta, G. Learn, A. Rodrigo, C. Rinaldo, M. Gorry, J. Mullins, P. Nara, and G. Ehrlich (1998). "Evolution of human immunodeficiency virus type 1 envelope sequences in infected invididuals with differing disease progression profiles," *Virology*, 241, 251–259.

Shankarappa, R., J. Margolick, S. Gange, A. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. Rinaldo, G. Learn, X. He, X. Huang, and J. Mullins (1999). "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection," *Journal of Virology*, 73, 10489–10502.

Shannon, C. (1948). "A mathematical theory of communication," *Bell System Technical Journal*, 27, 379–423.

Sommen, C., D. Commenges, S. Le Vu, L. Meyer, and A. Alioum (2010). "Estimation of the distribution of infection times using longitudinal serological markers of HIV: Implications for the estimation of HIV incidence," *Biometrics*, 67, 467–475.

StataCorp, ed. (2013). *Stata 13 Survey Data Reference Manual*, College Station, TX: StataCorp LP.

Stover, J. (2004). "Projecting the demographic consequences of adult HIV prevalence trends: the Spectrum projection package," *Sexually Transmitted Infections*, 80, i14–i18.

Stover, J., P. Johnson, T. Hallett, M. Marston, R. Becquet, and I. Timaeus (2010). "The Spectrum projection package: improvements in estimating incidence by age and sex,

mother-to-child transmission, HIV progression in children and double orphans," *Sexually Transmitted Infections*, 86, ii16–ii21.

Sukasih, A. and D. Jang (2005). "An application of confidence interval methods for small proportions in the health care survey of DoD beneficiaries," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3608–3612.

Suligoi, B., M. Massi, C. Galli, M. Sciandra, F. Di Sora, P. Pezzotti, O. Recchia, F. Montella, A. Sinicco, and G. Rezza (2003). "Identifying recent HIV infections using the avidity index and an automated enzyme immunoassay," *Journal of Acquired Immune Deficiency Syndromes*, 32, 424–428.

Taffe, P., M. May, and the Swiss HIV Cohort Study (2008). "A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort," *Statistics in Medicine*, 27, 4835–4853.

Templeton, A., M. Kramer, J. Jarvis, J. Kowalski, S. Gange, M. Schneider, Q. Shao, G. Zhang, M. Yeh, H. Tsai, H. Zhang, and R. Markham (2009). "Multiple-infection and recombination in HIV-1 within a longitudinal cohort of women," *Retrovirology*, 6, 54.

Towler, W., M. James, S. Ray, L. Wang, D. Donnell, A. Mwatha, L. Guay, C. Nakabiito, P. Musoke, J. Jackson, and S. Eshleman (2010). "Analysis of HIV diversity using a high-resolution melting assay," *AIDS Research and Human Retroviruses*, 26, 913–918.

Tweya, H., C. Feldacker, J. Estill, A. Jahn, W. Ng'ambi, A. Ben-Smith, O. Keiser, M. Bokosi, M. Egger, C. Speight, J. Gumulira, and S. Phiri (2013). "Are they really lost? 'True' status and reasons for treatment discontinuation among HIV infected patients on antiretroviral therapy considered lost to follow up in urban Malawi," *PLoS ONE*, 8, e75761.

Ukoumunne, O. (2002). "A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials," *Statistics in Medicine*, 21, 3757–3774.

UNAIDS (2011). "Global AIDS response progress reporting: monitoring the 2011 political declaration on HIV/AIDS: guidelines on construction of core indicators: 2012 reporting," Geneva, Switzerland.

UNAIDS Reference Group on Estimates Modeling and Projections (2005). "Statement on the use of the BED assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring," Geneva, Switzerland.

Wang, R. and S. Lagakos (2009). "On the use of adjusted cross-sectional estimators of HIV incidence," *Journal of Acquired Immune Deficiency Syndromes*, 52, 538–547.

Wang, R. and S. Lagakos (2010). "Augmented cross-sectional prevalence testing for estimating HIV incidence," *Biometrics*, 66, 864–874.

Wang, Y., S. Ray, O. Laeyendecker, J. Ticehurst, and D. Thomas (1998). "Assessment of hepatitis c virus sequence complexity by electrophoretic mobilities of both single- and double-stranded DNAs," *Journal of Clinical Microbiology*, 36, 2982–2989.

Wei, X., S. Ghosh, M. Taylor, V. Johnson, E. Emini, P. Deutsch, J. Lifson, S. Bonhoeffer, M. Nowak, B. Hahn, M. Saag, and G. Shaw (1995). "Viral dynamics in human immunodeficiency virus type 1 infection," *Nature*, 373, 117–122.

Wei, X., X. Liu, T. Dobbs, D. Kuehl, J. Nkengasong, D. Hu, and B. Parekh (2010). "Development of two avidity-based assays to detect recent HIV type 1 seroconversion using a multisubtype gp41 recombinant protein," *AIDS Research and Human Retroviruses*, 26, 61–71.

Welte, A., T. McWalter, and T. Bärnighausen (2009). "Reply to 'Should biomarker estimates of HIV incidence be adjusted?'," *AIDS*, 23, 2062–2063.

Welte, A., T. McWalter, O. Laeyendecker, and T. Hallett (2010). "Using tests for recent infection to estimate incidence: problems and prospects for HIV," *Euro Surveillance*, 15, pii: 19589.

West, B., P. Berglund, and S. Heeringa (2008). "A closer examination of subpopulation analysis of complex-sample suvey data," *The Stata Journal*, 8, 520–531.

Wilson, E. (1927). "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, 22, 209–212.

Wilson, E., W. Shao, J. Brooks, R. Dewar, M. Kearney, C. Rehm, T. Rehman, S. Kottilil, J. Coffin, and F. Maldarelli (2011). "New bioinformatic algorithm to identify recent HIV-1 infection," in *Proceedings of 18th Conference on Retroviruses and Opportunistic Infections*, Boston, MA, USA, abstract #1057.

Wilson, K., E. Johnson, H. Croom, K. Richards, L. Doughty, P. Cunningham, B. Kemp, B. Branson, and E. Dax (2004). "Incidence immunoassay for distinguishing recent from established HIV-1 infection in therapy-naive populations," *AIDS*, 18, 2253–2259.

Wolter, K. (2007). *Introduction to Variance Estimation*, New York: Springer, 2 edition.

World Health Organization (2009a). "Guidelines for surveillance of drug resistance in tuberculosis," Geneva, Switzerland.

World Health Organization, ed. (2009b). *WHO Technical Working Group on Statistical Approaches for Development, Validation and Use of HIV Incidence Assays*, Geneva, Switzerland, available at http://www.who.int/diagnostics_laboratory/links/hiviwg_geneva_04_09_report.pdf.

World Health Organization (2012a). "WHO HIV drug resistance report 2012," Geneva, Switzerland.

World Health Organization (2012b). "World Health Organization global strategy for the surveillance and monitoring of HIV drug resistance," Geneva, Switzerland.

World Health Organization (2012c). "World Health Organization protocol for population-based monitoring of HIV drug resistance emerging during treatment and related program factors at sentinel antiretroviral therapy clinics, 2012 update,," Geneva, Switzerland.

Wu, C. (1985). "Variance estimation for the combined ratio and combined regression estimators," *Journal of the Royal Statistical Society, Series B*, 47, 147–154.

Yamaguchi, Y. and T. Gojobori (1997). "Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts," *Proceedings of the National Academy of Sciences*, 94, 1264–1269.

Yansaneh, I. (2005). "Chapter II: Overview of sample design issues for household surveys in developing and transition countries," in United Nations Department of Economic and Social Affairs, Statistics Division, ed., *Household Sample Surveys in Developing and Transition Countries*, New York: United Nations, 11–34.

Young, C., D. Hu, R. Byers, S. Vanichseni, N. Young, R. Nelson, P. Mock, K. Choopanya, R. Janssen, T. Mastro, and J. Mei (2003). "Evaluation of a sensitive/less sensitive testing algorithm using the bioMèrieux Vironostika-LS assay for detecting recent HIV-1 subtype B' or E infection in Thailand," *AIDS Research and Human Retroviruses*, 19, 481–486.